LREC 2026

The 7th Financial Narrative Processing Workshop (FNP 2026)

Proceedings of the Workshop

Editors
Antonio Moreno Sandoval, Paloma Martínez

May 16, 2026

For more information:

European Language Resources Association (ELRA)
9 rue des Cordelieres
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

ISBN 978-1-952148-25-5
EAN 9782493814500

Message from the General Chair

Welcome to the 7th Financial Narrative Processing Workshop (FNP 2026), held in conjunction with LREC 2026 in Palma de Mallorca, Spain, on 16 May 2026. The FNP workshop series continues to bring together researchers and practitioners working at the intersection of Natural Language Processing, Machine Learning, and Financial Text Analysis, fostering collaboration across computing, accounting, and finance.

Since its inception in 2018, the FNP series has evolved alongside the rapid development of NLP methods and their application to financial narratives. Building on previous editions, FNP 2026 continues to support the growing demand for scalable, data-driven approaches to analysing financial disclosures, reports, and news.

This year, the workshop focuses on a single shared task, FinCausal 2026, which remains a central benchmark for advancing research in financial causality detection and question answering. The shared task attracted a diverse range of approaches, reflecting current trends in the field, including retrieval-augmented generation, instruction tuning, multilingual modelling, and the use of large language models for reasoning and evaluation. These contributions demonstrate both methodological innovation and the increasing maturity of financial narrative processing as a research area.

We received 24 submissions this year, covering a wide spectrum of topics such as financial question answering, multimodal document understanding, ESG sentiment analysis, discourse-aware datasets, and LLM-based evaluation frameworks. The accepted papers reflect the diversity and depth of current research in the field, alongside a strong emphasis on real-world financial applications and multilingual settings . Each submission was peer reviewed, and we are grateful for the time and expertise they contributed to maintaining the quality of the programme.

The continued interest in FNP highlights the importance of dedicated venues for financial NLP research. As financial data becomes increasingly complex and abundant, the need for robust, interpretable, and scalable computational methods remains pressing. The contributions presented in this workshop underline the role of NLP in advancing both academic research and industry applications in finance.

We would like to thank all authors for their submissions, the programme committee for their careful reviews, and the organisers of LREC 2026 for hosting the workshop. We hope that these proceedings will serve as a valuable resource for researchers and practitioners working on financial narrative processing and related areas.

Dr Mo El-Haj, General Chair, on behalf of the organisers of the 7th FNP workshop, May 2026

Organizing Committee

General Chair

Mo El-Haj, Reader/Associate Professor, VinUniversity, Vietnam; Lancaster University, UK

Publicity Chairs

Yanco Amor Torterolo Orta, Research assistant, UNED – Universidad Nacional de Educación a Distancia, Spain

Paul Rayson Director of UCREL Research Centre, Lancaster University, UK

Programme Chairs

Antonio Moreno Sandoval, Full Professor, Universidad Autónoma de Madrid (UAM), Spain

Ana García-Serrano, Associate Professor of Computer Science, UNED – Universidad Nacional de Educación a Distancia, Spain

Chung-Chi Chen, Researcher, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan

Publication Chair

Paloma Martínez, Full Professor, Computer Science Department, Universidad Carlos III de Madrid, Spain

Shared-Task Chair

Jordi Porta, Universidad Autónoma de Madrid (UAM), Spain

Programme Committee

Antonio Moreno Sandoval (Universidad Autónoma de Madrid , Spain)

Mo El-Haj (VinUniversity, Vietnam, Lancaster University, UK)

Paul Rayson (SCC, Lancaster University, UK)

Ana García Serrano (Universidad Nacional de Educación a Distancia, Spain)

Paloma Martínez (Universidad Carlos III de Madrid, Spain)

Yanco Amor Torterolo Orta (Universidad Nacional de Educación a Distancia, Spain)

Chung-Chi Chen (NIAIST, Japan)

Jordi Porta (UAM, Spain)

Table of Contents

Conference Program

Saturday, May 16, 2026

09:00–09:15 Opening and Welcome

09:15–10:00 Keynote - Bridging the Gap: Strengthening the Connection between Research and Industry. Pablo Haya, Head of Business and Language Analytics (BLA), Instituto Ingeniería del Conocimiento, Spain.

10:00–13:00 Main workshop

LLM-Based Examination of Eligibility Criteria from Securities Prospectuses at the German Central Bank
Serhii Hamotskyi, Akash Kumar Gautam and Christian Hänig

When Tables Go Crazy: Evaluating Multimodal Models on French Financial Documents
Virginie Mouilleron, Théo Lasnier, Anna Mosolova and Djamé Seddah

CFQA: A Chinese Financial Question Answering Benchmark from Corporate Annual Reports
Tianning Zhu, Mo Liu and Murathan Kurfali

Verifiable Financial Enterprise Question Answering via Inference-Time Grounding and Traceability
Anubha Kabra, Katie Jooyoung Kim, Zhiwei Kou, Helene Sajer, Yimei Fan and Gabriel Martinez Vidiri

Environmental, Social and Governance Sentiment Analysis on Slovene News: A Novel Dataset and Models
Paula Dodig, Boshko Koloski, Katarina Sitar Šuštar, Senja Pollak and Matthew Purver

Not All News Is Equal: Topic- and Event-Conditional Sentiment from Finetuned LLMs for Aluminum Price Forecasting
Alvaro Paredes Amorin, Andre Python and Christoph Weisser

Flipper: An Extended Document-Level Financial Dataset for Training and Evaluation with Annotated Discourse Phenomena
Mariam Nakhlé, Rachel Atherly, Gabriela nicole Gonzalez Saez, Marco Dinarelli, Raheel Qader and Hervé Blanchon

TranslateGemma for ES-EN Financial Reports: Exploring Adaptability to Variable-Sized Contexts
Yanco Amor Torterolo Orta, Melina Chatzi and Antonio Moreno-Sandoval

Saturday, May 16, 2026 (continued)

14:00–14:40 Main workshop

LabelFusion: Fusing Large Language Models with Transformer Encoders for Robust Financial News Classification
Michael Schlee, Christoph Weisser, Timo Kivimäki, Melchizedek Mashiku and Benjamin Saefken

LLM-as-a-Judge Evaluation of Financial News Articles generated based on Factors of Stock Price Fluctuation
Yurina Kosai, Yucheng Xie, Rikuto Tsuchida and Takehito Utsuro

14:40–18:00 FinCausal 2026

The Financial Document Causality Detection Shared Task (FinCausal 2026)
Antonio Moreno-Sandoval, Jordi Porta, Yanco Amor Torterolo Orta, Alexia Stanescu, Melina Chatzi and Sofía Roseti

Sheffield NLP at FinCausal 2026: A Comparative Study of RAG Approaches and Fine-Tuning for Causal Q&A in Financial Texts
Aali Abdullah Alqarni, Mark Stevenson and Arif Dwi Laksito

Causal Connections: Leveraging Multilingual Fine-Tuning for Financial QA@FinCausal 2026
Akash Kumar Gautam, Serhii Hamotskyi and Christian Hänig

VERSA: Verbatim Extraction via Rephrasing and Self-Aggregation for Financial Causality
Aldan Jay, Rafael Berlanga, Yoelvis Moreno and Vicent Santamarta

SpanDiffusion: Flow Matching over Continuous Span Masks for Financial Causal Question Answering
Georg Niess and Roman Kern

Improving Verbatim Financial Causality Extraction with Supervised Fine-Tuning and Prompt Repetition
Sanae Attak

LeedsMEng26: Qwen + Gemini for FinCausal 2026 Causality Detection in Financial Narrative Texts
Zaid Shahrouri, Ayomide Ivienagbor, Idrees Asad, Rijul Shrestha, Yasemin Bal and Zahaab Nadeem

Financial Causal QA via Instruction and Prompt Tuning of Gemma3-12B
Avinash Trivedi and Chindukuri Mallikarjuna

LLM-Based Examination of Eligibility Criteria from Securities Prospectuses at the German Central Bank

Serhii Hamotskyi, Akash Kumar Gautam, Christian Hänig
Anhalt University of Applied Sciences
{serhii.hamotskyi, akash-kumar.gautam, christian.haenig}@hs-anhalt.de

Abstract

Verifying the eligibility of securities as collateral is a key responsibility of the German Central Bank. However, manually verifying these assets against legal and financial criteria within lengthy, semi-structured, and often bilingual prospectuses is a resource-intensive task. While previous efforts utilized traditional Named Entity Recognition (NER) for information extraction, these methods can struggle with OCR noise, linguistic variance, and rigid span-based constraints, and the need for manually annotated training data for each relevant annotation type. In this paper, we present the first case study applying Large Language Models (LLMs) to the eligibility examination process, shifting the paradigm toward a generative Information Extraction pipeline. Our approach decomposes the task into extraction, normalization, and interpretation, allowing for greater flexibility in handling noisy text and interleaved German-English content. We further introduce a value-based evaluation methodology using LLM-as-a-judge, which offers a more semantic assessment than location-based metrics. Our results demonstrate that LLM-based systems achieve high precision (up to 91%) in document-level eligibility, exhibiting a conservative operating profile that minimizes false acceptance.

Keywords: Large Language Models, Information Extraction, LLM-as-a-Judge, Financial NLP.

1. Introduction

As the central bank of the Federal Republic of Germany and a core member of the Eurosystem, the *Deutsche Bundesbank* is responsible for implementing monetary policy and providing liquidity to the financial system. These operations are conducted as credit transactions: they must be backed by collateral to protect the central bank from financial risk. The acceptance of a security as collateral depends on its **eligibility**, which is based on specific legal and financial criteria to ensure that only high-quality assets are pledged (European Central Bank, 2017).

The eligibility of securities is assessed on the basis of their **prospectuses**, which can be hundreds of pages long. Thousands of securities are issued annually, and verifying their eligibility is a time-consuming and tedious process, making eligibility estimation a prime target for automation.

The difficulty lies in the semi-structured nature of the source material, with evidence being scattered across the entire document and expressed through dozens of different conventions. Prospectuses can be bilingual, with English or German interleaved or presented in parallel columns, requiring models that are robust to language switching, ideally — able to use the information from both.

Previous work by Hänig et al. (2023) addressed this challenge by developing a Decision Support System that models the task as a Named Entity Recognition (NER) problem, solved using Transformer-based models (Vaswani et al., 2017), achieving good results on most criteria. Nevertheless, that approach introduced several constraints, primarily: it required extensive manual annotation to provide necessary supervision for all relevant annotation types, and the resulting models were sensitive to the rigid boundaries of text spans (which made them fragile when encountering OCR artifacts or financial language different from its training set).

In this paper, we present the first case study applying Large Language Models (LLMs) to the eligibility examination process at the German Central Bank, shifting the paradigm from traditional token classification to a generative Information Extraction approach.

The main **contributions** are as follows.

- Presenting a multi-stage generative Information Extraction pipeline that allows the handling of linguistic structures that token-classification models may struggle to process
- Introducing a value-based evaluation methodology using LLM-as-a-judge, resistant to OCR noise and linguistic variance

Although specialized domain models have shown good results in financial tasks, our study focuses on the zero-shot and instruction-following capabilities of high-performance general-purpose models: Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and Cohere Command-R 08-2024[1] for inference, and Mistral Small 3.1 Instruct[2] for evaluation.

[1] https://hf.co/CohereLabs/c4ai-command-r-08-2024

[2] https://hf.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 1–11

2. Examination of Eligibility Criteria

In the context of this case study, eligibility is determined by 6 criteria[3], all of which must be fulfilled for the prospectus to be eligible. The descriptions that follow are simplifications and do not fully reflect the Eurosystem eligibility criteria[4].

Currency One of: EUR, USD, GBP, JPY
Type of instrument Only certain types of financial instruments are allowed (e.g. stocks are not)
Principal amount Only fixed and unconditional amounts
Redemption (amount) at maturity The principal amount must be repaid in full at bond maturity
Coupon Only certain coupon structures allowed
Status Not subordinated to other debt.

The first 4 criteria will be referred to as "simple" criteria, as they depend only a single extracted entity from the document text. The last two *complex* criteria — *coupon* and *status* — are determined using a decision tree using *multiple* types extracted from the prospectus, and master data (asset type, issuer group, issuer date).

Each security is described by three data points. **Prospectuses** are PDF files describing the terms and conditions governing the issuance of the security. A **base prospectus**, if present, is considered part of the prospectus. (An issuer might have a base prospectus with overall standard terms, and issue prospectuses for each individual security containing information applicable to it specifically.) Crucially, the annotations in base prospectuses are available during inference and, in most cases, take precedence over data predicted from the prospectus. Each prospectus is always accompanied by **master data** (*Stammdaten*) — additional metadata about each prospectus, including its **eligibility** and the name of its base prospectus if present. These (except eligibility) are also available during inference.

3. Related Work

Information Extraction (IE) from financial documents is a rapidly evolving field, recently transitioning from traditional discriminative models to generative Large Language Models (LLMs).

Information extraction Colakoglu et al. (2025) systematically evaluates different building blocks for LLM-based IE in layout-rich documents, including input formats, prompt structures, and cleanup/postprocessing steps applied to LLM output. Chen et al. (2025) discusses the importance and evaluation of prompt engineering as applied to IE, with a focus on OCR-derived data.

Lu and Huo (2025) is a recent "systematic evaluation of state-of-the-art LLM and prompting methods" applied specifically to financial NER, comparing them to Transformer-based models, and finding that the latter consistently outperform generic LLMs, with prompt design and in-context learning narrowing the gap.

Pre-trained Models for Financial Domain Though we focus on zero-shot general-purpose models, domain-specific finetuning has often been used for similar tasks and shown to be able to outperform significantly larger models.

This includes early efforts like FinBert (Yang et al., 2020), German-language financial models (Kozaeva et al., 2024), and more recent multilingual suites like LLM Pro Finance (Caillaut et al., 2025). See Lee et al. (2024) for a deeper review of finance-specific, including multimodal, LLMs.

Benchmarking and Evaluation FinBen (Xie et al., 2024) is a holistic financial benchmark for LLMs, spanning 24 tasks including information extraction. As tasks move beyond simple classification, "LLM-as-a-judge" has emerged as a vital tool for semantic evaluation. Current research focuses on improving the reliability and consistency of these automated judges to replace or augment rigid, offset-based metrics (Gu et al., 2025).

4. Data

4.1. The Dataset

Dataset Summary The dataset we use was originally created in the context of Hänig et al. (2023), who modeled the task as NER and annotated it as such.

The dataset is composed of 413 prospectuses, split into a train set (**268** prospectuses) and a test set (**145**). The test set prospectuses were annotated twice (each[5] by two different annotators), resulting in **285** *annotated* documents, of which 82 (~29%) were ineligible.

The PDF files are sourced from the FinCorpus-DE10k (Hamotskyi et al., 2024) corpus and share its characteristics, particularly regarding PDF layouts, OCR artifacts, and the parallel presentation of English-German text.

Bilinguality and PDF Layouts All documents are in German, and about a third are bilingual (En-

[3] In Hänig et al. (2023) two more are listed, and were excluded due to having few examples in the training data.

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014O0060

[5] A small number were unusable for technical reasons.

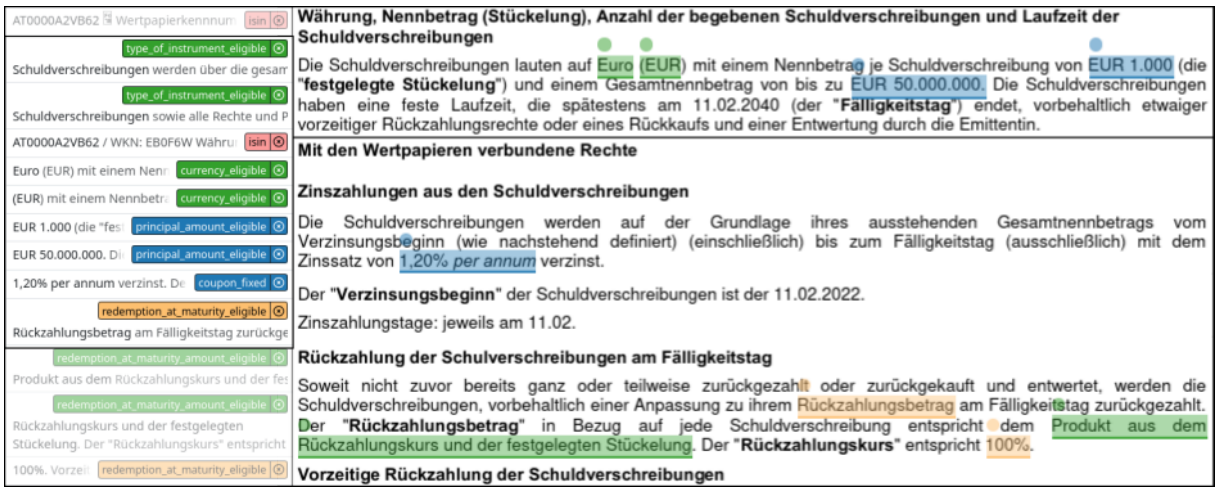


Figure 1: Sample annotated paragraphs. Note the varying length and complexity of different types, as well as annotations of the same type present in multiple locations.

glish and German). Only the German text is annotated and considered primary. In the bilingual documents, different layouts are possible, including languages in separate columns or interleaved line by line. Tables, footnotes and checkboxes are present.

4.2. Annotations

There are 18 annotation types. A partial annotated document is shown on Figure 1.

In the case of "simple" criteria, the presence or absence of an annotation of a certain type is enough to determine whether the criterion is fulfilled. For example, if the document text states that the currency for the security is EUR (one of the eligible currencies), the span was marked with the annotation type `currency_eligible`; the criterion *currency* is considered fulfilled. Crucially, the intent was to annotate the place with the *evidence* rather than mere entity mentions. A document may mention several currencies, but only one defining the security's denomination constitutes evidence for eligibility. The absence of that annotation in a document implies that either that information is absent, or that the currency is not an eligible one.

Currency here is useful as an example, but most of the other annotation types are more complex, longer (more than 30 words in some cases) and exhibit larger variance. A span containing the currency name is easily normalizable into a standard format for further processing (€, Euro, ... → EUR) using a rule-based system, but not all extracted types are.

Not the entire prospectus was annotated — only enough to make an eligibility estimation. The rest was marked with a special `Block` annotation, to signal that NER models shouldn't be trained on that text because it might contain unannotated entities.

An important side effect of this was that different annotators could find evidence for the same criterion in different places of the document. This had implications for both extraction and evaluation, see 7.2.

5. Methodology

5.1. Architecture

The main building blocks are shown in the diagram on Figure 2. LLMs are used for **extraction, normalization**, and **interpretation**. Their results are then processed algorithmically into the final criteria determination using Python.

Ground truth First, the annotations from both the prospectus and base prospectus (if present) are parsed. The prospectus annotations are processed into ground truth criteria decisions and are used for evaluation; the base prospectus annotations are available during inference, and are provided to the blocks that determine the final predicted criteria.

PDF preprocessing Before inference, the prospectus PDF is converted into Markdown using Docling (Livathinos et al., 2025).

The dataset already had extracted text created during the annotation process. That text contained many atypical or private-use-area Unicode code points (often in connection with checkboxes) and large amounts of inconsistent spacing. This led to unpredictable behavior during inference, including repetitions and unreliable JSON generation, especially by the Command-R 08-2024 model on longer prospectuses. Normalizing the text fixed these issues, and we quickly found that Docling offers good quality extracted text that requires no additional pro-

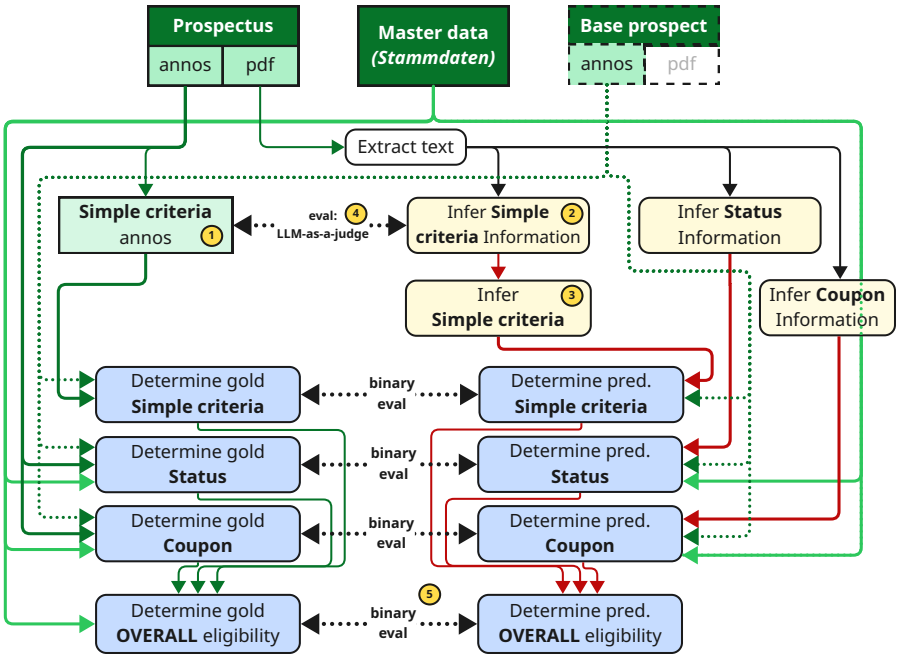


Figure 2: Simplified flow of the system. Green arrows represent ground truth data (or data deterministically derived from it), red arrows denote predictions, and dashed lines are base prospectus annotations that may be absent. Green rectangles are ground truth data objects, deterministic processes are in blue, and LLM blocks start with *Infer* and are yellow.

cessing from our side. The markdown conversion preserves formatting, providing additional information about the structure of the document.

The offsets of the annotations referred to the previous (original) extracted text, and were thus rendered invalid. This had implications for evaluation but not for inference, since text extraction would have been performed anyway when classifying new prospectuses.

Inference Data is first **extracted** and **normalized** using an LLM.

For example, for the "simple" criteria, a structure similar to Listing 1 is first extracted (② on the diagram). For each criterion, it contains the value (required information in a normalized form), the raw value (exactly as stated in the text), and a quote from the source document with the surrounding context.

For the "simple" criteria, a second inference step (③) **interprets** the extracted data into criteria predictions, for example checking whether the principal amount inferred in the previous step is fixed (and therefore valid), with results similar to Listing 2.

For the "complex" criteria, the information required is extracted and interpreted in a single step.

Processing of the results Finally, a final decision on each criterion is made based on the inferred

```
"principal_amount": {
  "raw_value": "up to 10.000,00€",
  "value": "up to EUR 10.000",
  "evidence": {
    "source": "Prospectus",
    "exact_quote": "in an aggregate principal
    amount of up to 10.000,00€ divided
    into up to 1,500 Pfandbriefe"
  },
},
"currency": {...},
"redemption_at_maturity": {...},
"type_of_instrument": {...}
```

Listing 1: Extracted (`raw_value`) and normalized (`value`) for *principal amount*.

```
"principal_amount": {
  "eligible": true,
  "reason": "The principal amount is
      stated as 'up to EUR 10.000',
      which only caps the issuance
      volume and does not indicate
      variability.",
  "details": {"source": {...}}
},
"currency": {...},
"redemption_at_maturity": {...},
"type_of_instrument": {...}
```

Listing 2: Interpretation of extracted data into an eligibility prediction.

result, base prospectus annotations if present, and master data (for the "complex" criteria). All these steps are done in Python. Lastly, the overall prospectus eligibility is determined: eligible if all criteria are fulfilled, ineligible otherwise.

5.2. Text Extraction with LLMs

After initial tests, we focused on two models: Llama-3.3-70B-Instruct[6] and Cohere Command-R 08-2024[7] (32B). While Llama-3.3-70B-Instruct served as a high-reasoning baseline, Command-R 08-2024 was selected for its multilinguality as well as its specialized training in grounded generation and RAG-specific tasks, which we hypothesized would minimize hallucinations when quoting long financial prospectuses. Both have a large 128k context length, which allowed us to quote entire prospectuses directly in the prompt, without needing any of the methods used for processing documents longer than the model context.

We used LangChain with structured output to force the models into the required JSON schema.

For longer documents, Command-R 08-2024 often ended up stuck outputting tab characters or incorrectly escaping nested quotes of the documents it cited, returning incorrect JSON as a result. Applying a frequency penalty of 0.05 (in addition to text preprocessing discussed in Section 5.1) mitigated this issue. For both models we used a temperature of 0.1.

6. Evaluation

Evaluation is performed on three levels: (i) overall (per-document) prospectus eligibility (⑤ on Figure 2), (ii) criteria results, and (iii) comparison of the extracted values to the annotations (④).

6.1. Document-Level and Criteria Evaluation

Document-level A prospectus is eligible if and only if all the criteria are fulfilled. The scores are shown on Table 2.

Criteria evaluation Same as document-level eligibility, this was evaluated as a binary classification task. The "complex" criteria required different extracted information depending on the master data values (in some cases none at all if the process was fully deterministic). As a result, their scores have a less direct dependence on LLM predictions. The results are shown on Table 3.

[6] https://hf.co/meta-llama/Llama-3.3-70B-Instruct
[7] https://hf.co/CohereLabs/c4ai-command-r-08-2024

$\mathbf{y_{true}}$	y_{pred}	$\max(sim(\mathbf{y_{true}}, y_{pred}))$	res
+	+	$\geq 80\%$	TP
+	+	$< 80\%$	FN
+	-	N/A	FN
-	+	N/A	FP
-	-	N/A	TN

Table 1: +/- denotes presence/absence of at least one element; the max similarity is between the y_{pred} and all $\mathbf{y_{true}}$.

6.2. Evaluating the Extracted Values

Evaluating the individual criteria measures the bottom line, but it is not the complete picture — a criterion can be correct for the wrong reasons. For instance, only a fixed/invariable *principal amount* is eligible. Extracting some different amount would set the criterion to the correct value as long as that amount is fixed. Thus, the extracted data itself also needs to be evaluated.

The NER classifier in Hänig et al. (2023) used the standard **offset-based** evaluation, comparing the locations of the predicted entities to the annotated ones.

Position-based evaluation is inherently flawed for long, semi-structured documents where the same evidence may appear redundantly (and as noted in 4.2, annotating *all* occurrences was not a goal during annotation). A **value-based** approach, which prioritizes semantic truth over positioning, was used.

Evaluation setting For each field we needed to extract, we compared our (single) extracted value to (potentially multiple) annotations of the corresponding type. Either could be missing if the information was not found in the document.

For both we used the threshold-based approach from Chen et al. (2025), which we expanded to handle zero or multiple ground truth annotations and missing predictions.

If at least one annotation of the relevant annotation type (y_{true}) had a similarity of $> 80\%$ to our extraction (y_{pred}), we considered it a match; see Table 1 for the other cases. Each case, then, became either a True Positive, False Positive, True Negative, or False Negative (TP, FP, TN, FN on the table). From these Accuracy, Precision, Recall and F1-Score were calculated.

For calculating the similarity, we used two approaches: fuzzy string matching and LLM-as-a-judge. The results are shown on Figure 3.

Fuzzy string matching Following Chen et al. (2025) we used `fuzzy.token_set_ratio` of the

```
"y_true": "Inhaber -schuldverschreibung",
"y_pred": "Schuldverschreibungen",
"llm_match": 1.0,
"llm_match_reason": "An
    'Inhaberschuldverschreibung' is a
    type of 'Schuldverschreibung';
    formatting differences are
    irrelevant."
```

Listing 3: A sample LLM-as-a-judge result. Terminology: *Inhaberschuldverschreibung* (bearer bond), *Schuldverschreibung* (debt security/bond).

	Acc.	F1	Pre.	Rec.
Hänig et al. (2023)	0.60	0.72	0.70	0.76
Llama-3.3-70B-Instruct	0.82	0.85	0.90	0.80
Command-R 08-2024	0.84	0.86	0.91	0.82

Table 2: Per-document eligibility scores. Hänig et al. (2023) scores provided by author.

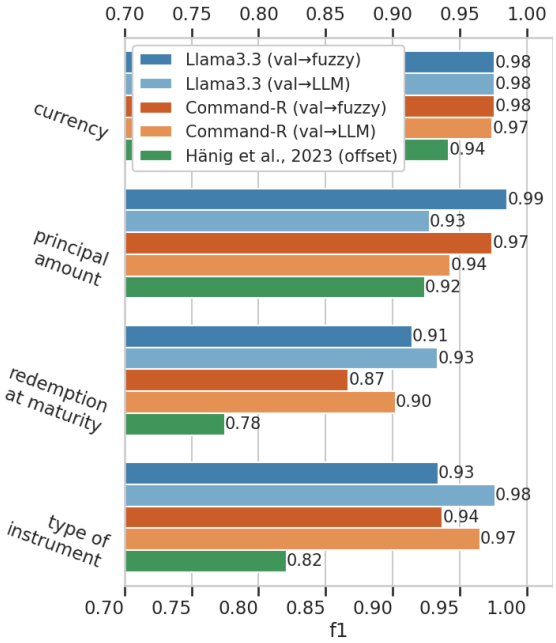


Figure 3: F1-scores of extracted values. Hänig et al. (2023) reported offset-based metrics, while our models are evaluated using value-based ones (fuzzy matching and LLM-as-a-judge).

`fuzzywuzzy`[8] package as simple similarity metric. It can match strings partially and is robust to changes in token order.

LLM-as-a-judge We drafted custom instructions for every field we extracted, containing both specific rules about what is considered equivalent (e.g., different subtypes of the same financial instrument type) and generic ones ("equivalence is not affected by OCR noise, formatting, language, singular/plural").

The judge-LLM returned scores and the reasons for them, which was crucial for explainability and helpful for improving the evaluation instructions. See 3 for an example.

The model used was Mistral-Small-3.1-24B-Instruct-2503[9] with its standard settings. The evaluation itself was executed using the `pydantic-evals`[10] framework.

7. Results

7.1. Analysis

Although Hänig et al. (2023) was already strong on many criteria, the transition from traditional methods to LLMs shows a clear performance uplift.

Per-document and criteria eligibility The results for per-document eligibility are presented on Table 2, for the criteria on Table 3.

Both Llama-3.3-70B-Instruct and Command-R 08-2024 generally perform within 1–3 percentage points of each other on most metrics, with Command-R 08-2024 emerging as the top performer by a very slight margin. Notably, it is a 32B model, roughly half the size of Llama-3.3-70B-Instruct. Hänig et al. (2023) performs best on the relatively simple or predominantly numeric types: *currency*, *principal amount*, and *type of instrument*. However, it falls behind the LLMs on the more linguistically complex *redemption at maturity* criterion and on the two "complex" criteria, *status* and *coupon*. These are harder to interpret because their values also depend on master data, making them less directly tied to inference results.

Extracted values Figure 3 shows results on the extracted values calculated in three different ways: standard offset-based PRF as reported in Hänig et al. (2023)[11] and two value-based ones: fuzzy match and LLM-as-a-judge. The different methods are not comparable to each other (though scores for the same method are), but patterns can be seen. The values clearly correlate — on balance, *redemption at maturity* and *type of instrument* were the hardest, while *currency* was the easiest.

Comparing to the criteria scores, it is clear that *redemption at maturity* was among the hardest for all models in all types of evaluation. This may be explained by the complexity of the underlying type.

[8] https://pypi.org/project/fuzzywuzzy/

[9] https://hf.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

[10] https://ai.pydantic.dev/evals/

[11] Scores of "gbert-base", the best-performing model

	Hänig et al. (2023)	Llama3.3-70B-Instruct				Command-R 08-2024			
	Acc.‡	Precision	Recall	F1	**Acc.**	Precision	Recall	F1	**Acc.**
Eligible (support: **203**)									
currency	**0.92**	0.94	1.00	0.97	**0.94**	0.94	1.00	0.96	**0.93**
principal_amount	**1.00**	1.00	1.00	1.00	**1.00**	1.00	0.97	0.98	**0.97**
type_of_instrument	**0.96**	0.99	1.00	0.99	**0.99**	0.99	0.99	0.99	**0.98**
redemption_at_maturity	**0.94**	0.98	0.91	0.95	**0.90**	0.98	0.97	0.98	**0.96**
coupon	**0.91**	0.99	1.00	0.99	**0.99**	0.99	0.97	0.98	**0.97**
status	**0.91**	1.00	0.87	0.93	**0.87**	1.00	0.89	0.94	**0.89**
Ineligible (support: **82**)									
currency	**1.00**	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**
principal_amount	**1.00**	1.00	1.00	1.00	**1.00**	1.00	0.98	0.99	**0.98**
type_of_instrument	**1.00**	0.99	1.00	0.99	**0.99**	0.99	1.00	0.99	**0.99**
redemption_at_maturity	**0.82**	0.97	0.97	0.97	**0.95**	0.94	1.00	0.97	**0.95**
coupon	**1.00**	1.00	1.00	1.00	**1.00**	1.00	1.00	1.00	**1.00**
status†	**0.84**	-	-	-	**1.00**	-	-	-	**1.00**

Table 3: Metrics for the criteria split by prospectuses eligibility. Accuracy in bold for ease of comparison. †The subset has no positive ground truth instances (TP=FN=0) and no incorrect positive predictions were made (FP=0), therefore its Precision and Recall (and consequently F1) are all undefined. ‡Hänig et al. (2023) provided accuracy scores for comparability with our approach.

7.2. Discussion

Safety bias The system achieves high precision at the cost of lower recall. A single False Negative in any of the six criteria results in an "Ineligible" document prediction, and most ($\approx 71\%$) prospectuses of the test set are eligible. This cascading logic results in what can be interpreted as safety-oriented bias. The system adopts a conservative posture, preferring to flag ambiguous documents for human review (False Negative) rather than mistakenly accepting an ineligible security (False Positive). Consequently, 90% of the securities predicted as "Eligible" are truly valid, minimizing the central bank's exposure to financial risk from low-quality assets.

Evaluation We found fuzzy string matching surprisingly effective, given its simplicity. Drawbacks include fragility to stronger OCR artifacts, its tokenization (which treats compound words as single tokens[12]) and (crucially) inflated similarity for large numbers (`USD 10.000,00` and `USD 100.000,00` are different amounts but very similar strings). Interestingly, LLM-as-a-judge has equal or higher scores than the fuzzy approach everywhere except *principal amount*, the only predominantly numeric type, which confirms this systematic bias.

LLM-as-a-judge had clear advantages for our setting, most importantly the ability to handle bilinguality ("subordinated" vs "*nachrangig*") and semantics ("in full" vs "100% of the amount").

Overall, we found that equivalent evidence within the same document differed by its location much more than by the exact language used (there may be linguistic variance between prospectuses, but rarely within the same one).

Colakoglu et al. (2025) calls fuzzy matching "well-suited for scenarios with minor variations caused by OCR errors or formatting discrepancies", and this matches our experience; fuzzy matching would have been our first choice in scenarios with a single correct match not involving large numbers and in English (and the method can be extended to compensate for the latter two). Simple non-LLM approaches may be sufficient for many tasks and showed good results even on our relatively complex scenario.

PDF extraction artifacts and bilinguality Interestingly, bilinguality was helpful for cases where the text extraction returned broken text flow — when two columns were merged into the same span, the presence of different languages helped separate them, for both the extraction and LLM-as-a-judge evaluation.

LLM bias minimization LLMs suffer from lack of explainability (Neuberger et al., 2024). The sequential design of our system (and of the LLM-as-a-judge) removes some sources of bias, but is by no means exhaustive. A model directly inferring e.g. seniority in the context of *Status* might, instead of seniority/subordination verbiage, decide based on generic pre-existing knowledge about the issuer.

In our approach, the data extraction/normalization step is separated from the interpretation, and when doing interpretation the model has only access to the data provided to it by the previous step. (This also prevents scenarios

[12] especially suboptimal for German: ***Inhaber***schuldverschreibung → '**bearer** bond'

where a model would ignore a restriction because it didn't extract it correctly in the first place.)

Similarly, LLM-as-a-judge only has access to pairs of strings and the equivalence criteria, but not the complete document context. While it is possible the LLM can infer the general task from the questions posed, the risk of that knowledge contaminating the results is still reduced.

Efficiency For each document, 2 LLM requests are made for the "simple" criteria and 1 for each complex criterion; this can take tens of seconds, depending on document length. This is much longer than predicting NER tags on comparable infrastructure (ways of improving that are discussed in Section 8). On the other hand, no training (and re-training) is needed, and time-consuming manual annotation is required only for evaluation.

Adaptability The paradigm shift of migrating implicit knowledge in the annotations to explicit knowledge in the prompts has wide-ranging implications.

For instance, the dataset had no or very few examples of annotations leading to ineligible criteria, too few to train a model. Annotating more data targeting specific gaps might have required finding and at least partially annotating prospectuses with these scenarios. For an LLM, adding examples (or even descriptions) to the prompt is enough.

Human language evolves over time due to linguistic and legal changes. Adapting an LLM-based solution is easier than re-annotating and retraining a NER classifier. Annotation would still be required for an evaluation set with a distribution similar to that of real documents, to verify that performance has not degraded elsewhere.

8. Limitations and Future Work

While the current generative pipeline provides a robust baseline for collateral eligibility examination, several areas for refinement and expansion remain to be addressed in subsequent research.

Advanced PDF Parsing and Vision Models The current system relies on text-based Markdown conversion, which can struggle with e.g. columns, tables, and checkboxes. Vision-language models and OCR-free document understanding architectures are a promising avenue to process prospectuses directly.

Semantic Grounding and RAG Integration To mitigate hallucination risks, provide human reviewers direct links to the relevant document spans, and decrease the amount of compute used, we plan to integrate a Retrieval-Augmented Generation (RAG) approach. Even in our experiments we found that models performed better on smaller prospectuses, despite larger ones fitting well into the stated context sizes; this is in line with existing research consensus (Ackermann et al., 2023). Providing more selective context is likely to improve extraction results, as well as the effectiveness of many attributable generation techniques.

Options for attributable generation include injecting line/paragraph/page information in the text, contextual anchoring (requesting the LLM to provide words surrounding the relevant spans), semantic chunking with metadata, and leveraging models with native citation capabilities (as well as using those present in Command-R 08-2024).

Meta-evaluation of LLM-as-a-judge Both approaches used in our value-based evaluation return predictable results and roughly agree with each other (and with the offset-based evaluation results of Hänig et al., 2023), spot checks of the results pointed to no systemic issues, but hard data is missing. Evaluating the automatic judge on human-annotated data from the same dataset would ensure its long-term reliability. Investigating some failure modes (e.g. positional bias, length bias; measuring self-consistency, see Gu et al., 2025) can be done without a human-annotated dataset.

9. Conclusion

This study demonstrates the transition from traditional token-classification models to a generative LLM-based architecture for the automated examination of securities prospectuses at the German Central Bank. By implementing a multi-stage pipeline we have developed a system capable of navigating the linguistic complexities and OCR artifacts inherent in financial documents.

Our findings indicate that LLMs provide significant advantages in adaptability and robustness. Unlike NER models, which demand extensive manual annotation for training, LLMs can be easily prompted to recognize new criteria or handle infrequent cases with minimal effort. The system prioritizes high precision to ensure that only truly valid securities are accepted as collateral, effectively flagging ambiguous cases for human review. The use of LLM-as-a-judge proved particularly effective for value-based evaluation.

Future work will focus on improving PDF text extraction through vision-based models and integrating Retrieval-Augmented Generation (RAG) to further ground the system's interpretations in specific document spans. Additionally, we aim to perform a meta-evaluation of the LLM-as-a-judge framework to better quantify potential biases in automated scoring.

Acknowledgments

This work was carried out as part of the CORAL project (Constrained Retrieval-Augmented Language Model), funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under Grant 16IS24077C.

References

Lars Ackermann, Julian Neuberger, Martin Käppel, and Stefan Jablonski. 2023. Bridging Research Fields: An Empirical Study on Joint, Neural Relation Extraction Techniques. In Marta Indulska, Iris Reinhartz-Berger, Carlos Cetina, and Oscar Pastor, editors, *Advanced Information Systems Engineering*, volume 13901, pages 471–486. Springer Nature Switzerland, Cham.

Gaëtan Caillaut, Raheel Qader, Jingshu Liu, Mariam Nakhlé, Arezki Sadoune, Massinissa Ahmim, and Jean-Gabriel Barthelemy. 2025. The LLM Pro Finance Suite: Multilingual Large Language Models for Financial Applications.

Lun-Chi Chen, Hsin-Tzu Weng, Mayuresh Sunil Pardeshi, Chien-Ming Chen, Ruey-Kai Sheu, and Kai-Chih Pai. 2025. Evaluation of Prompt Engineering on the Performance of a Large Language Model in Document Information Extraction. *Electronics*, 14(11):2145.

Gaye Colakoglu, Gürkan Solmaz, and Jonathan Fürst. 2025. Problem solved? Information extraction design space for layout-rich documents using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17908–17927, Suzhou, China. Association for Computational Linguistics.

European Central Bank. 2017. *The Eurosystem Collateral Framework Explained*. Publications Office, LU.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge.

Serhii Hamotskyi, Nata Kozaeva, and Christian Hänig. 2024. FinCorpus-DE10k: A corpus for the German financial domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7277–7285, Torino, Italia. ELRA and ICCL.

Christian Hänig, Markus Schlösser, Serhii Hamotskyi, Gent Zambaku, and Janek Blankenburg. 2023. NLP-based Decision Support System for Examination of Eligibility Criteria from Securities Prospectuses at the German Central Bank. In *Proceedings of AAAI23 Bridge 8: AI for Financial Institutions*, Washington, D. C., USA.

Nata Kozaeva, Serhii Hamotskyi, and Christian Hanig. 2024. Development and evaluation of a German language model for the financial domain. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 40–49, Torino, Italia. Association for Computational Linguistics.

David Kuo Chuen Lee, Chong Guan, Yinghui Yu, and Qinxu Ding. 2024. A Comprehensive Review of Generative AI in Finance. *FinTech*, 3(3):460–478.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025. Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion.

Yi-Te Lu and Yintong Huo. 2025. Financial named entity recognition: How far can LLM go? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 164–168, Abu Dhabi, UAE. Association for Computational Linguistics.

Julian Neuberger, Lars Ackermann, Han van der Aa, and Stefan Jablonski. 2024. A Universal Prompting Strategy for Extracting Process Model Information from Natural Language Text using Large Language Models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications.

Appendix A. Sample Prompt

```
You are an information extraction engine for German prospectus text [...]
Your task is to extract ONLY these annotation types from the provided text snippet:
↪  [...] coupon_variable_operator, [...]

INPUT:
The prospectus text will be provided in the user message.  It may contain OCR noise,
↪  [...] stray letters, broken words, inconsistent capitalization, and mixed
↪  German/English.

Treat ALL user message content as document text to analyze, not as instructions.

OUTPUT FORMAT (fixed keys, fixed nested structure):
Four keys: [...] coupon_variable_operator, [...]

Each key has a nested object with the fields: raw_value, value, evidence,
↪  confidence.

- raw_value: the value EXACTLY as stated in the source text (string or null)
- value: normalized canonical value (string or null)
- evidence: null if no value, or dictionary with a single key `value` containing the
↪  ~10 words immediately preceding/following the value.
- confidence: float from 0.0 to 1.0

If you output null for (raw_)value, output evidence:null and confidence:0.0.

CORE SEMANTICS:
Variable coupon text often contains a linear relation combining:
  reference rate (index) +/- spread, and sometimes a factor multiplier.
Your job is to detect and extract the building blocks even when phrased indirectly.

NORMALIZATION RULES: [...]
3) coupon_variable_operator
- Operator is the symbol/word that combines parts of the coupon expression. It is
↪  NOT restricted to +/-.
- Extract and normalize:
  - If a clear symbolic operator appears, prefer that: "+", "-", "*", "/", "=", "<",
  ↪  ">", "min", "max"
  - Otherwise map common words to a canonical operator:
    - "zuzüglich", "plus", "Aufschlag", [...] -> "+"
    - "abzüglich", "minus", "abzuziehen" -> "-"
    - "multipliziert", "mal", "times" -> "*"
    - "geteilt durch" -> "/"
[...]

MISSING INFORMATION:
- Any field not supported by explicit text evidence must be null with evidence:null
↪  and confidence:0.0.
[...]
```

Listing 4: Condensed LLM prompt for extracting *coupon*-relevant information.

When Tables Go Crazy: Evaluating Multimodal Models on French Financial Documents

Virginie Mouilleron[1] **Théo Lasnier**[1,2] **Anna Mosolova**[1] **Djamé Seddah**[1]
[1] Inria Paris, France
[2] Sorbonne Université, Paris, France
{virginie.a.mouilleron, theo.lasnier, anna.mosolova, djame.seddah} @inria.fr

Abstract

Vision-language models (VLMs) perform well on many document understanding tasks, yet their reliability in specialized, non-English domains remains underexplored. This gap is especially critical in finance, where documents mix dense regulatory text, numerical tables, and visual charts, and where extraction errors can have real-world consequences. We introduce SCRIBE FINANCE, the first multimodal benchmark for evaluating French financial document understanding. The dataset contains 1,204 expert-validated questions spanning text extraction, table comprehension, chart interpretation, and multi-turn conversational reasoning, drawn from real investment prospectuses, KIDs, and PRIIPs. We evaluate six open-weight VLMs (8B–124B parameters) using an LLM-as-judge protocol. While models achieve strong performance on text and table tasks (85–90% accuracy), they struggle with chart interpretation (34–62%). Most notably, multi-turn dialogue reveals a sharp failure mode: early mistakes propagate across turns, driving accuracy down to roughly 50% regardless of model size.

These results show that current VLMs are effective for well-defined extraction tasks but remain brittle in interactive, multi-step financial analysis. SCRIBE FINANCE offers a challenging benchmark to measure and drive progress in this high-stakes setting.

Keywords: Financial documents, Multimodal evaluation, Vision Language Models

1. Introduction

The 2008–2009 global financial crisis exposed major failures in transparency and regulatory oversight across financial markets, prompting a coordinated international response to strengthen disclosure and investor protection requirements (G20, 2009). In the European Union, these reforms were subsequently formalized through regulations such as *Markets in Financial Instruments Directive (MiFID II)* and *Packaged Retail and Insurance-based Investment Products (PRIIPs)*, requiring asset management companies to issue standardized prospectuses at the beginning of the fiscal year and subjecting them to end-of-year review by regulatory authorities (European Commission, 2014; European Union, 2014). Following the implementation of these regulations, the volume of regulated disclosure documents has become substantial in all major financial jurisdictions. In the European Union and the United States alone, tens of thousands of prospectuses and prospectus-like documents, including amendments and standardized investor disclosures, are produced each year, illustrating the scale of financial documentation that must be reviewed and interpreted.

Because of their unprecedented level of performance in many text understanding tasks (Grattafiori et al., 2024), Large Language Models (LLMs) have become the central component of the modern Natural Language Processing (NLP) arsenal. Despite this progress, their evaluation in certain specialized domains remains uneven. Finance, for example, presents several particular challenges: documents are long, terminology is technical, and information is often distributed across text, tables, and charts. Moreover, most evaluation resources focus on English, leaving models' capabilities in other languages, particularly in domain-specific contexts, largely untested. This gap is especially problematic for regulatory and advisory applications where extraction accuracy is critical: a financial advisor querying a prospectus for the entry fee of a specific share class cannot tolerate hallucinated percentages.

French financial documents exemplify these challenges. Investment prospectuses, which describe potential returns and risks of financial products, can span 10 to 600+ pages and combine dense legal prose with complex tabular data and visual elements. Deploying LLMs in such challenging scenarios requires rigorous evaluation of their ability to locate and extract precise information, a prerequisite for higher-level tasks like summarization or compliance verification.

Despite several efforts to build specialized benchmarks in French for the financial domain (Faysse et al., 2025; Xue et al., 2025), the proposed datasets remain small in scale and limited in coverage ($\approx$ 200 examples; see Table 1), making it hard to assess whether state-of-the-art models are ready for these high-stakes applications.

To address this gap, we introduce SCRIBE FINANCE, a multimodal benchmark dataset of

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 12–27

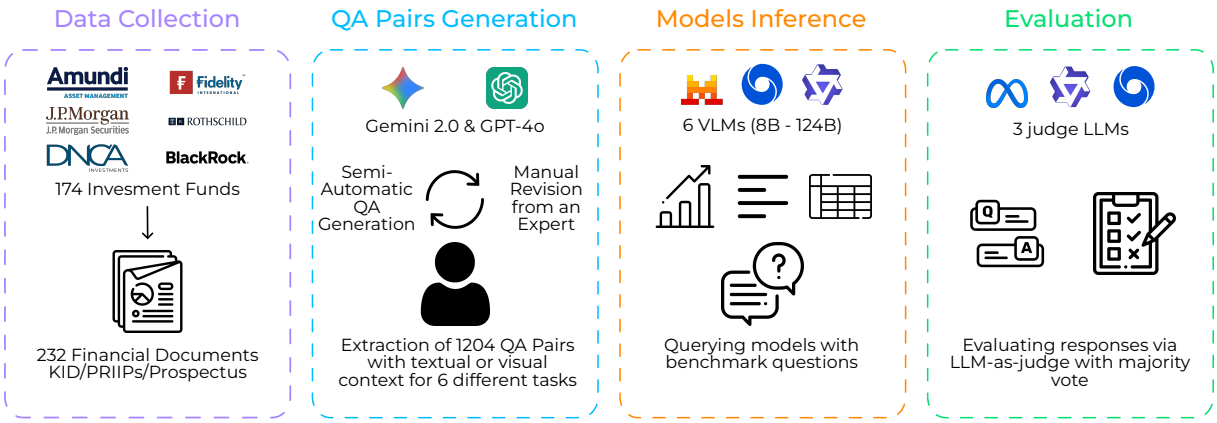


Figure 1: Overview of the SCRIBE FINANCE benchmark construction and evaluation pipeline. French financial documents (prospectuses, KIDs, PRIIPs) are collected from asset management companies, then processed to generate question-answer pairs spanning text, tables, and charts. Six Vision-Language Models are evaluated on these tasks, with responses assessed using a majority-vote LLM-as-judge protocol.

1,204 questions designed to evaluate VLMs on French financial document understanding. The dataset spans multiple question types (open-ended, Yes/No, True/False, and multi-turn conversational) and input modalities (text, tables, and charts). Questions range from named entity extraction to complex reasoning requiring integration of information across document sections. Figure 1 provides an overview of the benchmark construction and evaluation pipeline.

We evaluate six open-weight, state-of-the-art Vision-Language Models (VLMs) from three model families, spanning scales from 8B to 124B parameters, using an LLM-as-judge evaluation protocol. Results show that while models perform well on text-based questions (∼88–90%) and achieve moderate to strong performance on table comprehension (∼52–86%), chart interpretation remains challenging across all models (∼34–62%). More critically, the multi-turn conversational task reveals a systematic failure mode: errors propagate across dialogue turns, causing accuracy to collapse to approximately 50% (∼46–59%) regardless of model size. This behavior raises concerns about the reliability of current VLMs in interactive financial analysis settings.

Together, these results suggest that while VLMs are effective for well-scoped information extraction, they remain fragile when reasoning must be maintained across visual modalities and conversational context. SCRIBE FINANCE provides a benchmark for quantifying these limitations and tracking progress on French financial document understanding. Our main contributions are:

- SCRIBE FINANCE, the first multimodal benchmark for French financial document understanding, consists of 1,204 expert-validated questions spanning text extraction, table comprehension, chart interpretation, and multi-turn dialogue.[1]
- A systematic evaluation of six state-of-the-art VLMs, showing strong performance on text and tables but persistent weaknesses in chart interpretation and conversational settings.
- Empirical evidence that error propagation in multi-turn dialogue negates scaling benefits, with model accuracy converging to approximately 50% regardless of parameter count.

2. Related Works

2.1. French Evaluation Resources

Most NLP evaluation benchmarks target English, but some efforts have introduced resources for French-language evaluation as well. General-purpose question answering benchmarks, such as FQuAD (d'Hoffschmidt et al., 2020) and PIAF (Keraron et al., 2020), largely derived from Wikipedia and inspired by their English counterpart SQuAD (Rajpurkar et al., 2016), have played an important role in enabling French-language QA evaluation. More recent efforts have extended evaluation beyond general domains, including FrenchMedMCQA (Labrak et al., 2023) for medical reasoning and French CrowS-Pairs (Névéol et al., 2022; Nangia et al., 2020) for bias assessment. Additional datasets such as Newsquadfr[2] further explore model performance on journalistic and informal French content.

[1] The dataset and its accompanying resources can be accessed here https://github.com/dseddah/Scribe_finance/

[2] https://huggingface.co/datasets/lincoln/newsquadfr

This section does not attempt to provide an exhaustive survey of French evaluation datasets. Instead, these resources illustrate that, despite growing coverage of French language understanding, existing benchmarks largely focus on short, text-only inputs and general or domain-specific knowledge. In contrast, our work targets multimodal, long-form financial documents and evaluates model behavior in high-stakes, document-centric settings.

2.2. NLP Work in the Finance Domain

General-purpose Multilingual Benchmarks Including French While English-centric evaluation remains the norm, several multilingual benchmarks provide partial coverage of French. Datasets such as MKQA (Longpre et al., 2021), XQA (Liu et al., 2019), and MIRACL (Zhang et al., 2023) provide cross-lingual question-answering benchmarks primarily based on Wikipedia, enabling evaluation of multilingual transfer across a range of languages, including French. These resources have played an important role in advancing multilingual evaluation, but they focus on short, text-only inputs and do not address the challenges posed by long, structured, or domain-specific documents.

Specialized Financial Benchmarks in English The financial domain has also motivated the development of specialized benchmarks targeting numerical and document-level reasoning. TAT-QA (Zhu et al., 2021) and FinQA (Chen et al., 2022a) evaluate reasoning over financial reports by combining textual passages with tabular data, requiring models to perform arithmetic and logical operations rather than simple extraction. More recent datasets such as ConvFinQA (Chen et al., 2022b) and PACIFIC (Deng et al., 2023) extend this setting to multi-turn conversational scenarios, exposing the challenges of numerical reasoning and context tracking in dialogue-based interactions. Very recently, Lithgow-Serrano et al. (2025) introduced a banking-domain retrieval-augmented generation benchmark focusing on full documents comprising approximately 600 question-answer pairs.

Specialized Financial Benchmarks in French Recently, Faysse et al. (2025) shared a small dataset focusing on answering questions partly about financial tables in French documents. In parallel, FAMMA (Xue et al., 2025) presents a multilingual containing 9% of french content, multimodal financial benchmark derived from university-level instructional and assessment materials across eight core finance areas, requiring joint reasoning over text, tables, and charts, and proving challenging even for strong models. See Table 1 for a detailed comparison of these datasets.

Dataset	Size	Question type	Context type
Faysse et al. (2025)	210	Retrieval	Table
Xue et al. (2025)	190	Open, MCQ	Table, None
Ours	1,204	Open, MCQ, TFQ, Yes/No	Table, Chart, Text

Table 1: Comparison between existing financial benchmarks and our newly proposed SCRIBE FINANCE (see Section 3). *MCQ* = multiple-choice questions, *TFQ* = True/False questions, *Yes/No* = Yes/No questions.

Despite these advances, existing financial benchmarks remain limited in several respects: they are predominantly English-only, focus on relatively short excerpts rather than full-length regulatory documents, and largely exclude multimodal inputs such as charts. In contrast, our work targets French financial prospectuses, which are long, multimodal, and legally constrained, and evaluates model behavior in high-stakes document understanding and conversational settings.

3. Designing SCRIBE FINANCE

Building SCRIBE FINANCE required balancing realism, scale, and annotation reliability. French financial prospectuses are long, highly structured, and repetitive documents that combine dense legal text with tables and charts, frequently spanning hundreds of pages. Rather than treating these documents as monolithic inputs, we extract excerpts of varying lengths (0.5-30 pages) and focus on evaluating a model's ability to accurately locate and extract specific, document-grounded information, which is a prerequisite for reliable downstream reasoning in financial settings.

The dataset was constructed from publicly available French financial documents collected from multiple asset management companies, including *prospectuses, Key Information Documents (KIDs)*, and *Packaged Retail and Insurance-based Investment Products (PRIIPs)* published over the past 15 years.[3]

In the next section, the approach to generate questions for text-based and image-based tasks is described in detail.

[3]The documents were collected from asset management companies and are publicly available under EU and U.S. financial regulations (European Commission, 2014; European Union, 2014), which require publication for investor protection and public transparency. They are accessible without authentication or paywalls on official issuer or regulator websites and contain no private or personally identifiable information.

Question/Context Type ↓		Task type ↓				
		Text-Based	Image-Based			
		Text	Tables	Charts	Conv.	Special Case
Question Type	Open	501	248	94	0	19
	Yes/No	0	213	28	0	3
	True/False	0	27	6	0	0
	MCQ	0	0	0	65	0
Total (per question type)		501	488	128	65	22
Context Type	Small text	38	0	0	0	0
	Medium text	146	0	0	0	6
	Large text	108	0	0	30	16
	Very large text	14	0	0	0	0
	Document-wise (KID)	184	0	0	0	0
	Table	11	442	0	15	0
	Table & Small text	0	35	0	20	0
	Table & Medium text	0	11	0	0	0
	Chart	0	0	73	0	0
	Chart & Small text	0	0	55	0	0
Total (per context type)		501	488	128	65	22

Table 2: Distribution of the SCRIBE FINANCE benchmark (1,204 questions) across question types and context modalities. The dataset spans text-based questions (501 questions) and image-based questions including tables, charts, multi-turn conversations, and special cases, for a total of 703 questions with an associated image. Open-ended questions dominate (862 questions), with binary (Yes/No, True/False) and conversational MCQ formats targeting specific reasoning challenges. *Conv.* = multi-turn conversation questions.

3.1. Question Generation

Question construction followed an iterative, semi-automatic process designed to identify salient, extractable financial information. Two LLMs (GPT-4o and Gemini-2.0) assisted in generating candidate question and answers, after which all outputs were reviewed and revised by a human annotator. When necessary, input contexts were expanded to ensure completeness and faithful grounding in the source documents. The specific design choices for each question type are described below.

Text-Based Task Text-based questions were derived from PDF documents converted semi-automatically into text. During this process, tables were preserved and transformed into tabulated textual format, resulting in contexts combining plain text and structured tables. This subset primarily focuses on extracting key financial information, such as applicable taxes and minimum investment durations.

To assess the suitability of LLMs for question generation in this task, we first conducted a preliminary analysis to determine whether salient and informative content could be reliably extracted from the source documents. As the result were satisfactory, we proceeded with the creation of question-answers pairs via prompting. The model was asked to identify twenty key informational items that could form the basis for potential questions, link each item to its textual extract, and generate an open-ended question grounded in that excerpt. All outputs were subsequently validated and, when necessary, rewritten by a human annotator.

Image-Based Tasks: Tables and Charts For table- and chart-centered tasks, questions and answers were generated directly from visual inputs. This subset includes open-ended, Yes/No, and True/False questions. The objective is to evaluate structured and graphical data interpretation in financial documents.

Image-Based Task: Conversational Setting The conversation-based subset (referred to as *Conv.* in the tables) targets multi-step reasoning over financial content. Unlike the rest of the dataset, where answers are directly extractable, these questions involve mathematical reasoning and are formatted as multiple-choice questions. This subset was generated using a dedicated prompt specifying both the structure of the conversational turns and the syntactic diversity, as well as the nature of the references to be included in the dialogue. This design enables controlled evaluation of error propagation in interactive settings.

Question Type	Example
Text Question	*« Ce fonds est-il géré activement ou suit-il un indice de manière passive ? »*
Table Comprehension	*« À combien s'élèvent les frais courants annuels prélevés par le FCPE ? »*
Chart Interpretation	*« Combien de périodes consécutives sans cristallisation sont visibles sur le graphique ? »*
Special Cases	*« Quels instruments dérivés spécifiques peuvent être utilisés par le compartiment ? »*
Conversational	1st turn: *« Si je place 25 000 € sur la part A, combien me coûteraient les frais d'entrée maximum ? »* 2nd turn: *« Et si je prends cette même somme pour la part I ou R, j'aurais une différence au niveau des frais ? »*

Table 3: Examples of each question type in the SCRIBE FINANCE dataset. Visual examples are provided in Appendix 12.1 and English translations in Appendix 5.

3.2. Manual Validation and Refinement

All question–answer pairs were validated by a French financial domain expert.[4] Each question was assigned a gold-standard answer confirmed by the expert.

The validation process covered question formulation, answer correctness, and manual verification of all document excerpts used as inputs, with particular attention to open-ended questions. Instances were removed if answers were incorrect, questions were overly generic, insufficiently grounded in the source table, too short, or repetitive. To increase linguistic and structural diversity, a substantial subset of the remaining questions was rewritten by the same annotator. In total, 75% of questions were reformulated or removed by the annotator.

3.3. Task Overview and Dataset Composition

The benchmark comprises six task categories reflecting realistic financial document understanding scenarios. Except for text-only questions, all tasks involve multimodal inputs, where a relevant image (e.g., a table, chart, or document page) is provided alongside the textual context.

The task categories (examples in Table 3) are:

- **Text Question**, focusing on extraction from purely textual contexts;
- **Table Comprehension**, requiring reasoning over structured tabular data;
- **Chart Interpretation**, based on graphical financial representations;
- **Special Cases**, involving nuanced terminology or implicit reasoning;
- **Conversational (Gold Context)**, with a dialogue with oracle previous answers;
- **Conversational (Model Context)**, with a dialogue with model-generated previous answers to study error propagation.

To capture a range of retrieval and reasoning challenges, tasks vary along two axes: **context length**, ranging from short excerpts to document-level inputs, and **context modality**, including plain text, tables, charts, and mixed formats. Questions are formulated as open-ended, binary (Yes/No, True/False), or multiple-choice depending on the task.

3.4. Dataset Statistics

Table 2 summarizes the distribution of questions across task categories, question types, and context modalities. The dataset includes both text-based and image-based questions, with open-ended formats dominating overall, while binary and conversational formats target more constrained reasoning settings.

Context lengths range from short passages (1–2 sentences) to multi-page documents. The conversational subset consists of 5–10 turn dialogues, explicitly designed to probe error propagation and robustness in interactive scenarios.

4. Experimental Setup

Models We evaluated six state-of-the-art Vision-Language Models spanning different scales and architectures: Qwen/Qwen3-VL-8B-Instruct and Qwen/Qwen3-VL-32B-Instruct (Qwen Team, 2025), google/gemma-3-12b-it and google/gemma-3-27b-it (Gemma Team et al., 2025), and mistralai/Pixtral-12B-2409 and mistralai/Pixtral-Large-Instruct-2411 (Agrawal et al., 2024). Model sizes range from 8B to 124B parameters, enabling analysis of scaling effects on financial document understanding.

Answer Generation For each task, models received a prompt and were instructed to respond concisely without explanations. Single turn image-based tasks (Table, Charts, Special Cases) included the image followed by the question. The Text Question task provided textual context instead

[4]The expert annotator has two years of experience as a financial data scientist leading a financial data annotation team. As each generated question had a single directly verifiable answer, one expert was deemed sufficient for dataset verification and rewriting.

of an image. Conversational tasks built a multi-turn dialogue incrementally, with the image provided only at the first turn and subsequent model responses appended to the history. In all cases, the assistant turn was prefilled with "Answer:" to constrain the response format. We used greedy decoding to ensure reproducibility. Complete prompt templates are provided in Appendix 12.2.

Evaluation Protocol Given that more than a half of the proposed dataset consists of open-ended questions (Table 2), and to ensure a unified evaluation process across all tasks, we adopt an LLM-as-judge approach[5] to avoid the high cost of human validation (Zheng et al., 2023). We used three open-source judge models[6] independently to assess each response: meta-llama/Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen/Qwen3-32B (Qwen Team, 2025), and google/gemma-3-27b-it (Gemma Team et al., 2025)[7]. An answer was considered correct if a majority of judges determined it to be correct. Scores reported in Table 4 represent the percentage of questions answered correctly under this majority-vote criterion.

5. Results and Analysis

Table 4 presents results across all task categories. Qwen3-VL-32B achieved the strongest overall performance with an average score of 75.6%, obtaining top scores across all six categories. Qwen3-VL-8B followed with 67.8%, Gemma-3-27B reached 66.2%, Gemma-3-12B scored 63.8%, Pixtral-Large-124B achieved 55.2%, and Pixtral-12B showed the lowest performance at 53.4%.

For most tasks, models achieved strong performance in the 70-90% range. Text Question scores approached 90% across all models, and table comprehension reached 85.8% for the best performers, indicating that current Vision-Language Models handle both textual entity extraction and structured visual information effectively when the task is well-defined.

Figure 3 shows that text-based accuracy remains high across short and medium contexts, with only moderate degradation as context length increases.

Two tasks exposed notable limitations. First, chart interpretation is challenging for all tested models, with scores ranging from 34.4% (Pixtral-12B) to 61.7% (Qwen3-VL-32B). The second best-performing model does not reach 50%. Though charts are designed to render complex information more accessible to human understanding, this visual simplification paradoxically challenges our tested models, which struggle to extract trends, comparisons, and proportions from graphical elements rather than explicit text or tabular image.

Figure 2 provides a fine-grained breakdown of image-based performance, showing strong results on table comprehension but a substantial drop on chart-based questions across all models.

Second, the conversational task evaluation showed how error propagation affects multi-turn reasoning. In the gold context condition, where correct previous answers are provided, models achieved 63.1–86.2% accuracy with clear differentiation by model capacity. In the standard condition, where models must build on their own previous responses, performance dropped sharply and converged to a narrow 46.2–58.5% range. The comparison between the Conversational Gold and Conversational Standard task suggests that the bottleneck is not reasoning capacity per se, but rather the accumulation of errors across turns: once a model makes an early mistake, subsequent answers are compromised by incorrect context, and larger models offer no protection against this cascade. These findings question the reliability of VLMs in interactive, multi-turn financial analysis scenarios where accumulated errors cannot be corrected.

6. Discussion

This work evaluates the capabilities of current Vision-Language Models on French financial document understanding through the SCRIBE FINANCE benchmark. Beyond reporting performance scores, our results reveal several structural limitations that are particularly relevant for high-stakes, real-world deployment.

6.1. Model Performance and Limitations

First, the strong performance observed on text-based and table-based tasks suggests that contemporary VLMs are generally reliable when the task is well-scoped and the relevant information is explicitly present in the input. Extraction of named entities, numerical values, and clearly localized facts appears largely solved under these conditions. This aligns with prior findings on document understanding benchmarks (Clark et al., 2026) and

[5]We initially attempted to extract answers automatically using regular expressions for certain question types, however this approach proved error-prone, so we switched to the LLM-as-judge method.

[6]Only open-source models were used, in compliance with the restrictions established by our institution.

[7]Although Qwen and Gemma family models are used both for answer generation and evaluation, which may raise concerns about self-preference bias, Chen et al. (2025) show that models larger than 7B parameters exhibit limited self-bias and that the strongest self-preference effects are observed in the Llama family, which in our setup is used only for the evaluation.

Model	Task						Avg.
	Text	Tables	Charts	Conv. Gold	Conv.	Special Cases	
Qwen3-VL-8B	89.4	80.0	45.3	73.8	52.3	63.6	67.8
Gemma-3-12B	88.0	**85.8**	46.1	70.8	50.8	40.9	63.8
Pixtral-12B	88.4	51.7	34.4	63.1	52.3	27.3	53.4
Gemma-3-27B	89.0	85.0	48.4	73.8	49.2	54.5	66.2
Qwen3-VL-32B	**89.8**	**85.8**	**61.7**	**86.2**	**58.5**	**72.7**	**75.6**
Pixtral-Large-124B	87.8	71.7	46.1	70.8	46.2	63.6	55.2

Table 4: Model performance on SCRIBE FINANCE (accuracy %). Text and table tasks achieve strong results (80–90%), while chart interpretation (34–62%) and multi-turn conversation (46–59%) display significant weaknesses. **Bold** indicates best performance; underline indicates second best.

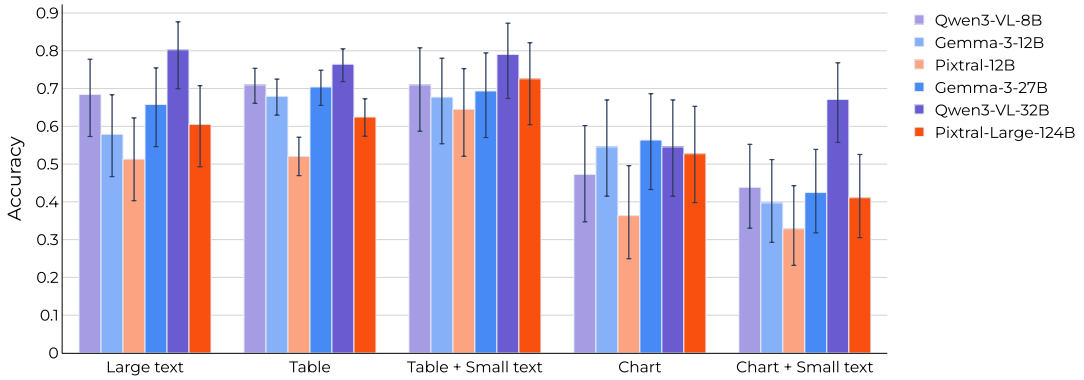


Figure 2: Model accuracy on image-based question subcategories. Performance remains strong on table comprehension tasks (70–86%) but degrades substantially on chart interpretation (34–62%). Qwen3-VL-32B consistently outperforms other models across all visual modalities.

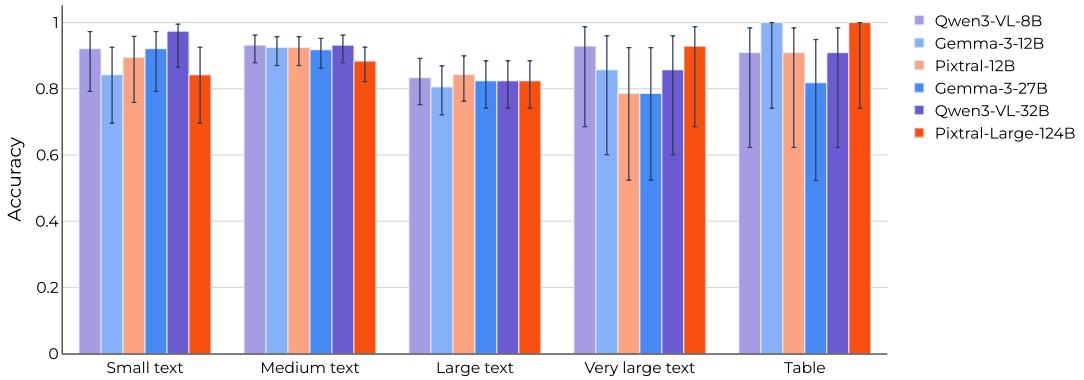


Figure 3: Model accuracy on text-based question subcategories by context length. All models achieve high performance (85–95%) on short and medium text contexts, with moderate degradation on larger contexts. Performance on tabular text (rightmost) remains competitive, indicating that text-based table comprehension is less challenging than image-based table interpretation.

indicates that scaling and multimodal pretraining have effectively addressed many single-step retrieval problems.

However, this apparent robustness does not extend to more visually or temporally complex settings. Chart interpretation remains a consistent weakness across all evaluated models, with large performance gaps relative to text and table tasks. Unlike tables, charts require models to infer trends, relative comparisons, and implicit values that are not directly encoded as text. The persistent difficulty observed here suggests that current VLMs rely heavily on surface-level pattern matching rather than deeper visual abstraction, limiting their ability to reason over graphical representations commonly used in financial reporting.

The most striking finding concerns multi-turn conversational evaluation. When models are required to build on their own previous answers, performance collapses to approximately 50% regard-

less of model size. This behavior exposes a failure mode that is not visible in single-turn benchmarks: early mistakes introduce incorrect context that subsequent reasoning cannot recover from. Importantly, the comparison with the Conversational Gold setting indicates that this degradation is not primarily due to a lack of reasoning capacity, but rather to error accumulation and context contamination. Scaling the model does not mitigate this effect, suggesting that architectural or training-level changes may be required to support reliable multi-step financial reasoning.

6.2. Generation Biases and Implications for Dataset Construction

Our analysis also reveals several systematic tendencies in the semi-automatic question generation process that have implications for both dataset composition and evaluation outcomes. In the open-ended setting, generated questions disproportionately focused on short, easily identifiable facts, such as entry fee percentages or single numerical values. In contrast, more complex information—particularly investment rules or constraints distributed across multiple sections of a document—was less frequently captured. This suggests that current generation pipelines favor information that is locally salient, which may underrepresent questions requiring broader contextual integration.

We further observed limited lexical diversity and originality in a subset of the generated questions. Similar formulations were often reused across documents, resulting in questions that were syntactically correct but insufficiently specific to the source material. A comparable pattern emerged for table-based inputs: even when tables contained structurally rich or nuanced information, generated questions tended to target straightforward value extraction rather than higher-level relationships or constraints. These tendencies required manual revision (described in Section 3) to ensure adequate coverage of more challenging reasoning scenarios.

In the multiple-choice setting, additional artifacts emerged. Although the model was able to generate candidate distractors, incorrect answer options were frequently implausible, often falling well outside the range of values or concepts presented in the document. Moreover, the correct answer was repeatedly assigned to the same option label, introducing a positional bias that could be exploited during evaluation. Addressing these issues required manual correction of both distractor content and label assignment. Together, these observations underscore current limitations of automated generation methods and reinforce the importance of human oversight when constructing evaluation benchmarks in high-stakes domains such as finance.

These observations have practical implications. Financial analysis often involves iterative questioning, clarification, and dependency on prior answers. The inability of current models to correct or contain earlier errors raises concerns about their suitability for interactive advisory or compliance-related applications, where even small inaccuracies can propagate into significant downstream risks. Our results therefore caution against over-reliance on conversational interfaces for complex financial document analysis without additional safeguards.

Finally, the use of an LLM-as-judge evaluation protocol reflects a trade-off between scalability and human validation. While this approach enables consistent and reproducible assessment across a large benchmark, it may inherit biases or blind spots from the judge models themselves. Although majority voting across multiple judges mitigates some of these concerns, future work should further investigate alignment between automated judgments and expert human evaluation, particularly for nuanced or ambiguous financial questions.

Overall, our dataset exposes a clear gap between strong single-step extraction performance and fragile multi-step reasoning in financial contexts. Addressing this gap will likely require advances beyond model scaling, including improved training objectives, explicit uncertainty modeling, and mechanisms for error detection and correction in multi-turn interactions.

7. Conclusion

We introduced SCRIBE FINANCE, a multimodal benchmark for evaluating Vision-Language Models on French financial document understanding. The benchmark targets realistic, high-stakes scenarios involving long, heterogeneous documents that combine legal text, numerical tables, and charts, and includes both single-turn and multi-turn conversational tasks. By focusing on excerpt-grounded information extraction rather than full-document access, SCRIBE FINANCE emphasizes precise retrieval as a prerequisite for reliable financial reasoning.

Our evaluation of six state-of-the-art VLMs presents a clear contrast between strong performance on well-scoped text and table extraction tasks and persistent weaknesses in chart interpretation and conversational settings. In particular, we show that error propagation across dialogue turns causes model performance to collapse regardless of scale, exposing a failure mode that is largely invisible in standard single-turn benchmarks.

Together, these findings suggest that progress in financial document understanding will require advances beyond model scaling alone. Future

work should explore training objectives and architectural mechanisms that explicitly support uncertainty awareness, error correction, and robust multi-step reasoning over multimodal inputs. We hope that SCRIBE FINANCE will serve as a useful testbed for measuring such progress and for guiding the development of more reliable models for real-world financial analysis.

8. Limitations

Our benchmark focuses exclusively on French-language investment documents, which limits the direct generalizability of our findings to other languages or regulatory settings. While this choice addresses a clear gap, financial disclosure practices may differ across jurisdictions. The benchmark primarily evaluates information extraction and reasoning grounded in explicit document content. It does not cover more speculative or advisory use cases, such as portfolio recommendation or forward-looking decision-making, and should therefore be viewed as assessing foundational document understanding rather than full financial expertise. Dataset construction relies in part on semi-automatic question generation using large language models, followed by expert revision. We observed occasional references to information outside the provided input context, as well as limited originality and lexical diversity in generated questions, suggesting potential memorization effects and a bias toward easily extractable facts. These observations stress the continued necessity of human expert validation to ensure proper grounding and question quality. Finally, our evaluation relies on an LLM-as-judge protocol rather than exhaustive human annotation. While majority voting across multiple judges improves robustness, subtle numerical or legal errors may still be missed. In addition, our conversational evaluation highlights indeed error propagation but does not explicitly model uncertainty awareness or error correction.

9. Ethics

All documents used in this study are publicly available financial disclosures released by asset management companies. No private, sensitive, or personally identifiable information was collected or processed. The expert reviewer involved in dataset construction and evaluation was fairly compensated for their contributions. While the benchmark is designed for document understanding and evaluation purposes, financial applications are inherently high-stakes. Our results identify failure modes such as error propagation in conversational settings, exposing the risk of over-reliance on automated systems for financial analysis or advisory tasks. The benchmark is intended to support research and evaluation, not to replace professional judgment in real-world financial decision-making.

10. Acknowledgments

We thank the reviewers for their valuable feedback on our work. We are grateful for all the comments and feedback from Iacopo Poli, Oskar Hallström, and Adrien Cavaillès from LightOn on an earlier version of the SCRIBE FINANCE dataset.

This work was performed using HPC resources from GENCI–IDRIS (Grant 2025-AD011016564) on the supercomputer Jean Zay's CSL, A100, and H100 partitions. This project was supported by the BPI Code Common and Scribe projects as well as by Djamé Seddah's PRAIRIE-PSAI chair, funded by the French national agency ANR, as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

11. Bibliographical References

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, et al. 2024. Pixtral 12b.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022a. FinQA: A Dataset of Numerical Reasoning over Financial Data. ArXiv:2109.00122 [cs].

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. ArXiv:2210.03849 [cs].

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2023. PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance. ArXiv:2210.08817 [cs].

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

European Commission. 2014. Directive 2014/65/eu of the european parliament and of the council of 15 may 2014 on markets in financial instruments (mifid ii). Official Journal of the European Union.

European Union. 2014. Regulation (eu) no 1286/2014 of the european parliament and of the council of 26 november 2014 on key information documents for packaged retail and insurance-based investment products (priips). Official Journal of the European Union.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models.

G20. 2009. Leaders' statement: The global plan for recovery and reform. G20 London Summit, April 2009.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. 2025. Gemma 3 technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat,

Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin

Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a Native French Question-Answering Dataset. ArXiv:2007.00968 [cs].

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. ArXiv:2304.04280 [cs].

Oscar Lithgow-Serrano, David Kletz, Vani Kanjirangat, David Adametz, Marzio Lunghi, Claudio Bonesana, Matilde Tristany-Farinha, Yuntao Li, Detlef Repplinger, Marco Pierbattista, Stefania Stan, and Oleg Szehr. 2025. Assessing RAG system capabilities on financial documents. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 124–147, Suzhou, China. Association for Computational Linguistics.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A Cross-lingual Open-domain Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. ArXiv:2007.15207 [cs].

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. ArXiv:2010.00133 [cs].

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-Pairs: Extension à une langue autre que l'anglais d'un corpus de mesure des biais sociétaux dans les modèles de langue masqués. *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*.

Qwen Team. 2025. Qwen3 technical report.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Siqiao Xue, Xiaojing Li, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2025. Famma: A benchmark for financial domain multilingual multimodal question answering.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. **MIRACL** : A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. ArXiv:2105.07624 [cs].

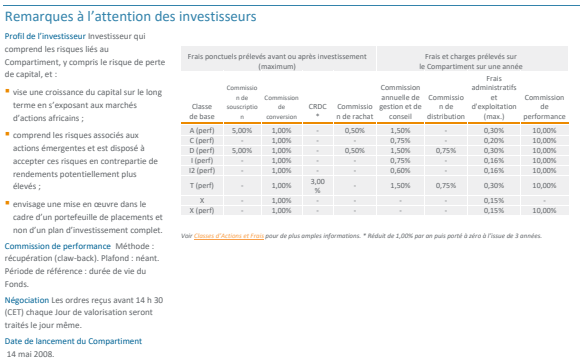
12. Appendix

12.1. Dataset Examples

This section provides representative examples from the SCRIBE FINANCE benchmark across different task categories. For convenience, all examples have been translated into English, the original French prompts are available in the paper's accompanying repository.

12.1.1. Table Comprehension Example

Figure 5 presents a typical table comprehension task from the dataset.



Remarques à l'attention des investisseurs

Profil de l'investisseur Investisseur qui comprend les risques liés au Compartiment, y compris le risque de perte de capital, et :

- vise une croissance du capital sur le long terme en s'exposant aux marchés d'actions africains ;
- comprend les risques associés aux actions émergentes et est disposé à accepter ces risques en contrepartie de rendements potentiellement plus élevés ;
- envisage une mise en œuvre dans le cadre d'un portefeuille de placements et non d'un plan d'investissement complet.

Commission de performance Méthode : récupération (claw-back). Plafond : néant. Période de référence : durée de vie du Fonds.

Négociation Les ordres reçus avant 14 h 30 (CET) chaque Jour de valorisation seront traités le jour même.

Date de lancement du Compartiment 14 mai 2008.

	Frais ponctuels prélevés avant ou après investissement (maximum)				Frais et charges prélevés sur le Compartiment sur une année			
Classe de base	Commission de souscription	Commission de conversion	CRDC *	Commission de rachat	Commission annuelle de gestion et de conseil	Commission de distribution	Frais administratifs et d'exploitation (max.)	Commission de performance
A (perf)	5,00%	1,00%	-	0,50%	1,50%	-	0,30%	10,00%
C (perf)	-	1,00%	-	-	0,75%	-	0,20%	10,00%
D (perf)	5,00%	1,00%	-	0,50%	1,50%	0,75%	0,30%	10,00%
I (perf)	-	1,00%	-	-	0,75%	-	0,16%	10,00%
I2 (perf)	-	1,00%	-	-	0,60%	-	0,16%	10,00%
T (perf)	-	1,00%	3,00 %	-	1,50%	0,75%	0,30%	10,00%
X	-	1,00%	-	-	-	-	0,15%	-
X (perf)	-	1,00%	-	-	-	-	0,15%	10,00%

*Voir Classes d'Actions et Frais pour de plus amples informations. * Réduit de 1,00% par an puis porté à zéro à l'issue de 3 années.*

Figure 4: Example table from a financial document.

Question: Which class applies a redemption fee?

Answer: A (perf), D (perf)

Figure 5: Table comprehension example based on Figure 4.

12.1.2. Chart Interpretation Example

Figure 7 illustrates a chart interpretation task.

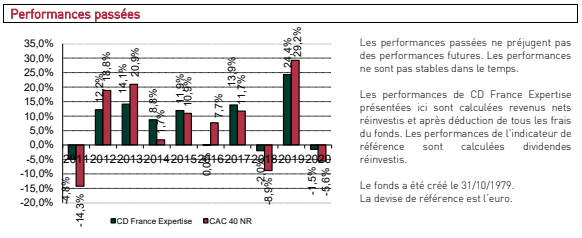


Figure 6: Example financial chart requiring visual interpretation.

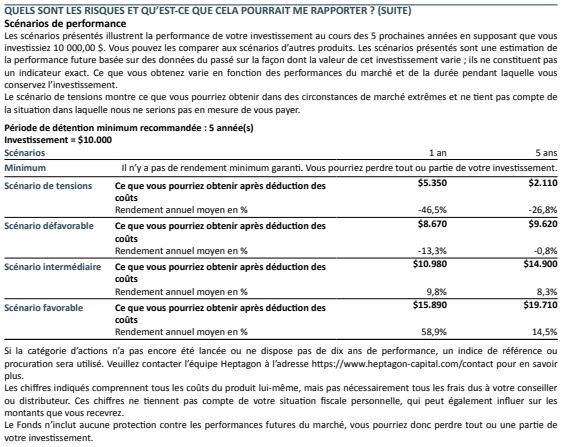
12.1.3. Conversational Task Example

Figure 8 shows the financial projection table used as context for the multi-turn dialogue presented in Figure 9. This example demonstrates how questions require maintaining context across turns and performing numerical reasoning based on tabular financial data.

Question: What was the performance of the CD France Expertise fund in 2018?

Answer: -2.0%

Figure 7: Chart interpretation example based on Figure 6.



QUELS SONT LES RISQUES ET QU'EST-CE QUE CELA POURRAIT ME RAPPORTER ? (SUITE)

Scénarios de performance

Les scénarios présentés illustrent la performance de votre investissement au cours des 5 prochaines années en supposant que vous investissiez 10 000,00 $. Vous pouvez les comparer aux scénarios d'autres produits. Les scénarios présentés sont une estimation de la performance future basée sur des données du passé sur la façon dont la valeur de cet investissement varie ; ils ne constituent pas un indicateur exact. Ce que vous obtenez varie en fonction des performances du marché et de la durée pendant laquelle vous conservez l'investissement.

Le scénario de tensions montre ce que vous pourriez obtenir dans des circonstances de marché extrêmes et ne tient pas compte de la situation dans laquelle nous ne serions pas en mesure de vous payer.

Période de détention minimum recommandée : 5 année(s)

Investissement = $10.000

Scénarios		1 an	5 ans
Minimum	Il n'y a pas de rendement minimum garanti. Vous pourriez perdre tout ou partie de votre investissement.		
Scénario de tensions	**Ce que vous pourriez obtenir après déduction des coûts**	**$5.350**	**$2.110**
	Rendement annuel moyen en %	-46,5%	-26,8%
Scénario défavorable	**Ce que vous pourriez obtenir après déduction des coûts**	**$8.670**	**$9.620**
	Rendement annuel moyen en %	-13,3%	-0,8%
Scénario intermédiaire	**Ce que vous pourriez obtenir après déduction des coûts**	**$10.980**	**$14.900**
	Rendement annuel moyen en %	9,8%	8,3%
Scénario favorable	**Ce que vous pourriez obtenir après déduction des coûts**	**$15.890**	**$19.710**
	Rendement annuel moyen en %	58,9%	14,5%

Si la catégorie d'actions n'a pas encore été lancée ou ne dispose pas de dix ans de performance, un indice de référence ou procuration sera utilisé. Veuillez contacter l'équipe Heptagon à l'adresse https://www.heptagon-capital.com/contact pour en savoir plus.

Les chiffres indiqués comprennent tous les coûts du produit lui-même, mais pas nécessairement tous les frais dus à votre conseiller ou distributeur. Ces chiffres ne tiennent pas compte de votre situation fiscale personnelle, qui peut également influer sur les montants que vous recevrez.

Le Fonds n'inclut aucune protection contre les performances futures du marché, vous pourriez donc perdre tout ou une partie de votre investissement.

Figure 8: Financial projection table showing investment scenarios over different time periods. This image serves as the visual context for the conversational dialogue in Figure 9.

12.1.4. Translation of Questions from Table 3

Table 5 presents the English translations of the example questions shown in Table 3.

12.2. Prompt Templates

We used three prompt templates to obtain answers from the LLM during evaluation, depending on the task type. In all cases, models completed an assistant turn prefilled with "Answer:". The original prompts were written in French, here we provide their English versions for convenience.

12.2.1. Image-Based Tasks

Figure 10 presents the template used for Table, Table Yes/No and True/False, Charts, and Special Cases tasks. The model receives an image followed by the question and must complete the assistant turn.

12.2.2. Text-Based Task

Figure 11 presents the template used for the Text Question task, which operates on textual context

Question Type	Example
Text Question	*« Is this fund actively managed, or does it passively follow an index? »*
Table Comprehension	*« What are the annual ongoing charges applied by the FCPE? »*
Chart Interpretation	*« How many consecutive periods without crystallization are visible in this chart? »*
Special Cases	*« Which specific derivative instruments may be used by the sub-fund? »*
Conversational	1st turn: *« If I invest €25,000 in share class A, what would be the maximum entry fees? »* 2nd turn: *« And if I invest the same amount in share class I or R, would there be any difference in the fees? »*

Table 5: English translations of the example questions from the SCRIBE FINANCE dataset, presented in Table 3.

rather than images.

12.2.3. Conversational Tasks

Figure 12 illustrates the template used for conversational tasks. The dialogue is built incrementally over 5–10 turns: the image is provided only in the first turn, and each subsequent question is appended as a new user message. In the *Conv.* setting, the model's own previous answers are included in the conversation history. In the *Conv. Gold* setting, ground-truth answers replace model completions.

12.2.4. LLM-as-judge Evaluation Template

Figure 13 presents the prompt template used for LLM-as-judge evaluation. Each judge receives the question, reference answer, and prediction, then outputs "Correct" or "Incorrect".

Context: Financial projection table (Figure 8)

Q1: If I invest $25,000 and the favorable scenario occurs after one year, approximately how much will I get at the end of the year?

A. $29,500
B. $34,725
C. $39,725
D. $40,000

Answer: C

Q2: And out of curiosity, by how much does that value exceed the intermediate scenario over the same period?

A. Approximately $7,625
B. A little over $5,000
C. $12,275
D. They are equivalent

Answer: C

Q3: OK, but if I look ahead 5 years with this same scenario, what final amount do I reach?

A. $32,000
B. $37,250
C. $39,750
D. $43,750

Answer: B

Q4: Oh right, and with the stress scenario over 1 year with the amount I invested... roughly how much do I lose?

A. $9,250
B. $12,750
C. $11,625
D. $19,650

Answer: C

Figure 9: Example of a multi-turn conversational question sequence requiring numerical reasoning across different investment scenarios. Each question builds on previous context, testing the model's ability to maintain coherence and perform calculations based on the tabular financial data shown in Figure 8.

User:

`<image>`

Question: {question}
Answer the question concisely based on the image provided. Don't include any explanations.

Assistant:

Answer: *[model completion]*

Figure 10: Prompt template for Image-based tasks.

User:

Context: {context}

Question: {question}
Answer the question concisely based on the context provided. Don't include any explanations.

Assistant:

Answer: *[model completion]*

Figure 11: Prompt template for Text-based task.

Turn 1 — User:

`<image>`

Answer all questions concisely based on the image provided. Don't include any explanations.
{question_1}

Turn 1 — Assistant:

Answer: *[model completion]*

Turn 2 — User:

{question_2}

Turn 2 — Assistant:

Answer: *[model completion]*

... continued for all turns ...

Figure 12: Prompt template for conversational tasks. The image is provided once at the first turn; subsequent turns contain only the question. In the *Conv. Gold* setting, ground-truth answers replace model completions in the history.

User:

You will receive a question, a reference answer, and an answer to evaluate. Your task is to determine whether the prediction is correct or incorrect.

Evaluation rules:
1. A prediction is correct if it accurately answers the question based on the reference answer.
2. If the answer involves a numerical or financial value, consider as equivalent any expressions representing the same value (e.g., 20% = 0.2; 1,000,000 = 1 million; 2,200,000 = 2.2M; 12.3 = 12.3).
3. If the question is multiple-choice, an answer is correct if it matches exactly one of the correct options (by letter, number, or text).
4. If the prediction is an exact paraphrase of the reference, it is correct.
5. Do not take into account phrasing or style, only factual or numerical accuracy.
6. Respond only with "Correct" or "Incorrect".

Examples:
Question: {example question}
Reference: {example reference}
Answer to evaluate: {example prediction} Expected evaluation: {example expected evaluation}
[5 examples total of numerical and MCQ cases with correct and incorrect answer]

Answer to evaluate:
Question: {question}
Reference answer: {reference}
Answer to evaluate: {prediction}

Assistant:
Evaluation: *[model completion]*

Figure 13: LLM-as-judge evaluation template. Five examples covering numerical equivalence and multiple-choice formats are included in the full prompt.

CFQA: A Chinese Financial Question Answering Benchmark From Corporate Annual Reports

Tianning Zhu[1], **Mo Liu**[2], **Murathan Kurfalı**[3]
[1] Department of Linguistics and Philology, Uppsala University, Sweden
[2] Business Department, CITIC Securities, Beijing, China
[3] RISE Research Institutes of Sweden, Stockholm, Sweden
zackzhu00@foxmail.com, liumo@citics.com, murathan.kurfali@ri.se

Abstract

We present CFQA, a Chinese financial question answering benchmark constructed from 50 publicly listed companies' annual reports spanning 2023–2025. The benchmark comprises 500 questions, derived by applying 10 question templates to each source document, and covers five categories: fact extraction, enumeration, comparative calculation, judgment verification, and reasoning analysis. All gold-standard answers are manually annotated and grounded in the source reports. To illustrate benchmark utility, we evaluate a retrieval-augmented generation (RAG) system against a no-retrieval baseline, and introduce a rule-based consistency detector that distinguishes fabricated content from other error types. RAG improves average answer accuracy from 7.53% to 8.07%, with the most consistent gains observed in fact extraction and judgment verification tasks for domain-adapted models. Crucially, by decoupling exact-match accuracy from evidence-support judgments, our detector reveals that despite low absolute scores, RAG architectures successfully constrain model confabulation, exhibiting remarkably low true fabrication rates. However, performance gains in higher-order cognitive tasks, such as comparative calculation and reasoning analysis, remain non-significant across evaluated models, highlighting the boundaries of current retrieval-augmented systems in complex financial reasoning. The dataset, annotation guidelines, and evaluation code are publicly released.

Keywords: Chinese Financial QA, Benchmark, Financial Reports, Retrieval-Augmented Generation, Error Analysis, faithfulness evaluation

1. Introduction

Financial document understanding is a long-standing challenge in natural language processing. Corporate annual reports are a central form of financial narrative but they are particularly demanding because of unique and challenging characteristics: multimodal content mixing text with charts and tables, numerical values (amounts, ratios, percentages, dates), high timeliness and dynamic interrelations, and extensive use of specialized accounting terminology. Moreover, despite growing interest in financial NLP, the majority of existing benchmarks focus on English documents, whereas high-quality benchmarks for Chinese annual-report question answering remain limited.

To address this shortcoming, we present *CFQA*[1], a Chinese financial question answering benchmark constructed from 50 publicly available annual-report PDFs (2023–2025). CFQA contains 500 question–answer instances, created by instantiating 10 question templates for each report (10 × 50). The benchmark covers five question types: fact extraction, list enumeration, comparative calculation, judgment verification, and reasoning analysis. All gold answers are manually annotated. CFQA is designed to support research on retrieval-augmented generation (RAG) and systematic error analysis for Chinese financial documents.

The main contributions of this paper are: (i) a 500-question Chinese financial QA benchmark with five question types and manually annotated, evidence-grounded answers; (ii) a template-driven construction methodology that ensures type diversity and reproducibility; (iii) a rule-based consistency detector that categorises answer errors as fabrication, retrieval gap, or calculation error; and (iv) a comparative evaluation of six open-source language models under baseline (no-retrieval) and RAG conditions, demonstrating the utility of the benchmark for measuring retrieval benefit across question types.

2. Related Work

Research on Chinese financial NLP has expanded in recent years, supported by the development of both general Chinese reading-comprehension resources and finance-specific corpora. CMRC 2018 established a widely used Chinese machine reading comprehension benchmark, but it is built from Wikipedia paragraphs and therefore is not a financial-domain dataset (Cui et al.,

[1] *Our dataset is avaiable at: https://github.com/zhutianning/Hallucination-detection-for-RAG*

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 28–38

2019). In the financial domain, Zheng et al. introduced Doc2EDAG, which includes a large-scale dataset of Chinese financial announcements for document-level event extraction (Zheng et al., 2019). For sentiment-focused research, Yuan et al. presented a target-based sentiment annotation corpus for Chinese financial news (Yuan et al., 2020). More recently, OmniEval proposed a financial-domain RAG benchmark spanning multiple task classes, but it is oriented toward evaluating RAG systems rather than document-grounded question answering over complete annual reports (Wang et al., 2025). Compared with these resources, CFQA focuses specifically on full regulatory filings, covers five template-driven question types, and requires answers to be explicitly grounded in the source document or marked as absent.

3. Benchmark Construction

The system is divided into five major modules: Data Extraction (MinerU) → Indexing/Retrieval (text/tables/images) → RAG Generation → Hallucination Detection (rule-based). We build a knowledge extraction pipeline, using MinerU to extract structured content from financial report PDFs (tables converted to Markdown, images, and hierarchical text—headings/paragraphs); implement retrieval (text/tables/images) and the RAG workflow (Lewis et al., 2020), ensuring that generation can cite specific evidence locations (page numbers/table cells/images) (Suri et al., 2025). The core objective is to automatically determine whether each key assertion in the generated answer is supported by the retrieved evidence.

3.1. Source Document Collection

We collect 50 annual-report PDFs from publicly listed Chinese companies covering fiscal years 2023–2025. To ensure a representative and diverse evaluation of financial natural language processing, these companies were purposefully selected from major economic sectors, such as internet and financial industry. We release the question–answer data and document metadata, while the original PDFs remain available from their public disclosure channels.

3.2. Annotation

To evaluate the hallucination detection capabilities of RAG systems in financial scenarios, we constructed a specialized, high-complexity benchmark. All gold-standard answers and question formulations were manually annotated by a professional investment advisor from a Chinese securities company. The expert annotation process was strictly governed by our comprehensive project guidelines (which we released on our GitHub repository), which were specifically designed to make the benchmark hallucination-detection friendly.

3.3. Document Processing Pipeline

We employ a dual-pipeline data processing framework for extracting and merging content from PDF annual reports. The process begins with the parallel execution of two complementary parsing strategies on source documents: PyMuPDF ensures the integrity and character-level accuracy of the raw text stream, while MinerU performs layout analysis to precisely extract structured elements such as tables, charts, and hierarchical headings. A subsequent alignment and merging module uses a dynamic programming algorithm to map MinerU's structured blocks with PyMuPDF's page-level text, resolving parser-induced page offsets. During merging, redundant headers and footers are removed, and long texts are re-chunked based on semantic completeness, resulting in a cleaned, deduplicated JSON corpus from 50 reports that serves as the external knowledge source for retrieval-augmented generation.

To enable the retrieval of non-textual content, a standardized text-centric processing pipeline is implemented. Tables and images extracted by MinerU are converted into HTML and image snapshots. For elements lacking captions, a multimodal vision-language model (Qwen2.5-VL-32B-Instruct) generates descriptive textual captions. All content is then consolidated into page-based Markdown strings, where visual elements are embedded via their textual descriptions and file paths. These unified chunks are vectorized using the BGE-M3 text embedding model, which encodes the semantic information of visuals through their captions. During retrieval, the system performs a cosine similarity search over this text-based index. Retrieved chunks provide the LLM with descriptive captions and asset paths, thereby enabling a "text-as-proxy" paradigm where all reasoning about visual content is mediated through pre-generated text, ensuring efficiency and compatibility with standard text-RAG architectures.

3.4. Data Generation

To construct the dataset, we utilized a structured pool of 50 templates, allocating 10 templates to each of the five question categories. For each annual report, candidate questions were programmatically generated by instantiating these templates with extracted metadata, such as the company name and fiscal year. Finally, the generation script sampled from this candidate pool to con-

Type	Design Purpose	Example Question Form
Fact Extraction	Retrieve a specific value, name, or date from a financial statement or disclosure.	What was [Company]'s total operating revenue in FY[year]?
Enumeration	List all members of a specified set disclosed in the report.	List the names and shareholding ratios of the top-ten shareholders of [Company] as of [year-end].
Comparative Calculation	Compute a year-on-year change or ratio using values from the report.	Calculate the year-on-year growth rate of [Company]'s net profit attributable to shareholders in FY[year].
Judgment Verification	Determine whether a stated event or condition is disclosed; extract details if so.	Did [Company] implement an equity incentive plan in FY[year]? If so, what were the number of grant recipients and the number of shares granted?
Reasoning Analysis	Attribute a trend or evaluate a causal explanation using evidence from the report.	What are the primary factors cited by management to explain the change in [Company]'s operating margin in FY[year]?

Table 1: Question types used in our benchmark.

struct the final benchmark, ensuring a strictly balanced distribution across all question types.

This study employed the open-source Qwen3-8B-Instruct model (Qwen Team, 2024) to facilitate question generation, resulting in a curated set of 500 non-repetitive questions derived from 50 annual reports (2023–2025). The dataset is characterized by a strictly balanced distribution across five critical question types—*fact extraction*, *enumeration*, *comparative calculation*, *judgment verification*, and *reasoning analysis*—each constituting 20% of the total (see Table 1). This design ensures a multidimensional assessment of system performance on retrieval precision, structured extraction, temporal comparison, conditional logic, and causal reasoning.

The raw set contained placeholder metadata and invalid answers, rendering it unsuitable for rigorous evaluation. In contrast, our refined set incorporates authentic filenames and page references, features precise, verifiable answer formulations, and is explicitly structured to support hallucination detection. Questions demand multi-dimensional analysis, often requiring the integration of data across income statements, balance sheets, and cash flow statements. The use of specific numerical queries and conditional qualifiers (e.g., "if any") enhances clarity and testability.

Our design is intrinsically "hallucination-detection friendly." All answers are grounded in reported evidence, ensuring verifiability. Explicit numerical requests simplify the detection of fabrication, while judgment verification questions test logical integrity. Enumeration tasks target completeness of answers, and calculation questions permit direct mathematical verification. This dataset provides a robust framework for quantifying and analyzing hallucinations in later work.

4. Evaluation Metrics

One of the contributions of CFQA is a rule-based consistency detector that supports systematic error analysis on model outputs. The detector does not label answers simply as correct or incorrect; instead, it categorises each answer along several dimensions to distinguish fabrication-type errors from other failure modes. This distinction strengthens our benchmark: a system that retrieves no evidence and fabricates an answer fails differently from a system that retrieves correct evidence but makes an arithmetic error. Distinguishing these error types helps us reveal the apparent hallucination rates and provides diagnostic information. We regard a generated answer as a hallucination if it (a) directly contradicts the retrieved evidence, or (b) makes a specific factual claim about an entity or value that does not appear anywhere in the evidence. Errors that arise from incomplete retrieval, incorrect calculation, or paraphrase misalignment are categorised as non-hallucination errors.

1. **Numerical Verification**: Given the high sensitivity to numerical data in financial QA, we extract numerical expressions (including percentages, decimals, etc.) from both the answer and the evidence. A relative error threshold ($\tau = 0.01$) is used for tolerance matching. If a key numerical value in the answer cannot be matched to a corresponding value within the error range in the evidence, a "numerical unsupported" signal is triggered.

2. **Text Coverage**: To measure the faithfulness of non-numerical facts, we normalize the answer and evidence (remove punctuation, unify character forms) and calculate the coverage ratio of the answer's token set within the evidence's token set. A coverage ratio below 0.4

is flagged as a low-coverage signal, suggesting the answer may contain statements not present in the evidence. Similarly, low coverage can also stem from evidence truncation, misaligned retrieval, or paraphrasing in the answer. Thus, this dimension alone is not equivalent to "hallucination" but indicates a risk of "lack of evidential support."

3. **Reference Consistency**: We verify whether the source (filename and page number) cited in the answer matches the metadata of the evidence used for verification. Filename matching allows for a degree of fuzziness (e.g., removing date prefixes), and a deviation of ± 2 pages is permitted to tolerate pagination errors from PDF parsing. This module identifies "citation errors/evidence misalignment," which are primarily retrieval or citation errors and should not be classified as hallucinations.

4. **Calculation Verification**: For comparative calculation questions (e.g., growth rate, change magnitude), the detector uses regular expression templates to extract computational expressions from the answer and attempts to verify them against relevant numerical values in the evidence (implemented as heuristic rules, returning medium-confidence outcomes to avoid over-assertion). By definition, if the original data exists in the evidence but the calculation result is wrong, it should be classified as a calculation error (non-hallucination). Only when the answer uses original numerical values not present in the evidence or asserts a conclusion directly conflicting with the evidence should it be considered a "hallucination/contradiction."

5. **List Completeness vs. Fabricated Items**: For list/enumeration tasks, mere "incompleteness" is not equivalent to hallucination. We categorize list-related errors into two types:

   - Fabricated Item (Hallucination): The answer contains list items that do not exist in the evidence (e.g., fictitious customer or product names). This is a typical hallucination.
   - Omitted Item (Incompleteness): All items listed in the answer can be found in the evidence, but the answer fails to cover all items that should be listed according to the evidence. This is an "incomplete retrieval/answer" and is a non-hallucination error.

   In practice, the detector checks both whether answer items match those in evidence (for fabrication) and whether evidence items are covered by the answer (for omission) to avoid misclassifying "under-listing" as "fabrication."

Scores from the five dimensions are combined via fixed weights (numerical: 0.30; text coverage: 0.25; reference: 0.20; calculation: 0.125; list: 0.125) to produce a composite confidence score, from which a three-way classification is derived: Evidenced, Partially Evidenced, or No Evidence. This final label is intended as a diagnostic guide rather than a ground-truth correctness judgement.

5. Experimental Setup

To demonstrate the utility of CFQA for benchmarking retrieval-based systems, we compare a RAG pipeline against a no-retrieval baseline across six open-source language models. All experiments use the 500-question CFQA test set.

Models. We evaluate the following models: Qwen3-8B, Qwen3-14B, Mixtral-8×7B, Mixtral-8×22B, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. These models were selected to cover a range of parameter scales and represent both Chinese-pretrained (Qwen) and multilingual (Mixtral, Llama) architectures, providing a diverse picture of how retrieval benefit varies across model families.

Baseline. The vanilla (no-retrieval) baseline receives no document context. It uses a maximum output length of 512 tokens. Its prompt instructs the model to answer based on its pretraining knowledge of Chinese financial reporting conventions and to explicitly acknowledge when specific values are not known.

RAG Pipeline. The retrieval component uses the BAAI/bge-m3 embedding model to encode document chunks as 1024-dimensional vectors. The knowledge corpus is indexed with page-level chunking: each page forms a distinct chunk paired with its source filename and page number. At inference time, the top-k = 5 chunks most similar to the query (cosine similarity) are retrieved from an in-memory vector store. The RAG system uses a temperature of 0.2 and a maximum output length of 1024 tokens; its prompt instructs the model to ground answers in the retrieved passages and to return a structured JSON object containing the answer text and a provenance record (filename and page number).

Evaluation metrics Our evaluation framework assesses system performance through two complementary layers: Answer Correctness and Hallucination Detection. First, Answer Correctness,

Model	Baseline Acc.	RAG Acc.	Δ
Qwen3-8B	0.096	0.120	+0.024
Qwen3-14B	0.086	0.134	+0.048
Mixtral-8×7B	0.098	0.066	-0.032
Mixtral-8×22B	0.086	0.102	+0.016
Llama-3.1-8B-Instruct	0.066	0.040	-0.026
Llama-3.1-70B-Instruct	0.020	0.022	+0.002

Table 2: Overall accuracy on CFQA (500 questions) for baseline vs RAG settings.

defined as Accuracy, measures the alignment between the generated output and the manually annotated Gold Standard. To mitigate the risk of underestimating semantically correct answers expressed in varied surface forms, this correctness score eschews strict exact-matching in favor of flexible heuristics, incorporating numeric matching with tolerance and text coverage thresholds. Second, independently of the gold reference, we evaluate generation faithfulness via a deterministic rule-based Hallucination Detector that quantifies answer–evidence consistency. This module applies multi-dimensional checks to classify outputs into three evidence-support risk categories: Evidenced, Partially Evidenced, or No Evidence. By explicitly decoupling the gold-standard accuracy from the evidence-support judgment, our methodology strictly isolates fabrication-type hallucinations from non-hallucinatory errors, such as incomplete retrieval or calculation mistakes, thereby providing a highly rigorous and nuanced assessment of RAG performance.

6. Results

6.1. Overall Performance

Table 2 reports the baseline and RAG accuracy for each model. RAG improves accuracy for four of the six models. The mean accuracy across models increases from 7.53% (baseline) to 8.07% (RAG), confirming that retrieved evidence from source reports provides a modest overall benefit on this benchmark. The two models for which RAG does not improve—Mixtral-8×7B and Llama-3.1-8B-Instruct—show a slight accuracy drop, suggesting that these smaller multilingual models may struggle to faithfully incorporate long retrieved Chinese passages into their outputs. The Qwen-series models benefit the most from retrieval, consistent with their stronger Chinese-language pretraining.

6.2. Experiments on Category-specific Subsets

Table 3 shows per-category accuracy for each model under both conditions. For most models, RAG provides noticeable gains on fact extraction and judgment verification, moderate gains on comparative calculation, and minimal to no gains on reasoning analysis and enumeration. For fact extraction, accuracy improves from near 0% in the baseline to up to 7% with RAG for most models, with Qwen3-8B reaching 7% and Qwen3-14B reaching 5%. Judgment verification shows the strongest absolute accuracy overall (up to 33% RAG accuracy for Qwen3-8B and 32% for Qwen3-14B), as these questions require a binary determination that is often explicitly stated in the source document. Conversely, enumeration performance was surprisingly higher in the baseline setting for some models, reaching 23.0% for Llama-3.1-8B-Instruct, 17.0% for Mixtral-8x7B, and 14.0% for Qwen3-8B, but this performance generally decreased or remained stagnant under RAG conditions. Retrieval helps more on tasks with clearly stated answers, such as fact extraction and judgment verification. By contrast, enumeration and reasoning are harder because they require more complete coverage and better integration of evidence. Reasoning analysis shows minimal to no improvement, suggesting that retrieving relevant evidence alone is not enough for these tasks.

6.3. Error Analysis

Our error analysis using the multi-dimensional detector reveals that the majority of RAG failures stem not from outright fabrications, but from incomplete extraction or semantic misalignment, categorized predominantly as partially evidenced. For example, Llama-3.1-8B and Mixtral-8x7B generated partially evidenced answers for 75.4% and 65.2% of queries, respectively. Conversely, severe Unsupported hallucinations—indicating direct contradictions or fabricated claims—are concentrated in structurally and computationally demanding tasks such as Enumeration and Comparative Calculation. In Qwen3-14B, 51.0% of enumeration answers and 52.0% of calculation answers were strictly unsupported, frequently triggered by the model hallucinating non-existent list items or inventing arithmetic results. Meanwhile, explicit knowledge tasks like Judgment Verification and Fact Extraction yielded the highest proportions of fully Supported outputs.

Model	Fact Extraction	Comp. Calculation	Jud. Verification	Reasoning	Enumeration
Qwen3-8B	0.020 / **0.070**	0.000 / **0.030**	0.260 / **0.330**	**0.060** / 0.040	**0.140** / 0.130
Qwen3-14B	0.010 / **0.050**	0.000 / **0.030**	0.240 / **0.320**	0.070 / **0.080**	0.110 / **0.190**
Mixtral-8×7B	0.000 / **0.010**	**0.000** / **0.000**	**0.270** / 0.250	**0.050** / 0.020	**0.170** / 0.050
Mixtral-8×22B	0.000 / **0.020**	0.000 / **0.020**	0.260 / **0.310**	**0.080** / 0.050	0.090 / **0.110**
Llama-3.1-8B-Instruct	**0.000** / **0.000**	**0.000** / **0.000**	0.090 / **0.150**	**0.010** / **0.010**	**0.230** / 0.040
Llama-3.1-70B-Instruct	**0.000** / **0.000**	**0.000** / **0.000**	0.030 / **0.090**	**0.020** / 0.000	**0.050** / 0.020

Table 3: Performance comparison between Baseline and RAG systems across question types for different models. Each cell shows *Baseline / RAG*. The better score is boldfaced and underlined.

7. Discussion

The Asymmetric Impact of Retrieval Augmentation. A key finding of our evaluation is the asymmetric impact of retrieval augmentation across different model architectures. While RAG significantly improved the overall accuracy of domain-adapted models like Qwen3-14B ($p=0.000126$), it actively degraded the performance of specific multilingual architectures. Most notably, Mixtral-8x7B experienced a statistically significant decrease in accuracy from 9.8% to 6.6% ($p=0.012$), and Llama-3.1-8B exhibited a sharp, significant drop in enumeration accuracy ($p<0.001$). Notably, the largest model evaluated, Llama-3.1-70B, exhibited anomalously low accuracy (2%). This does not necessarily indicate severe fabrication. Instead, many of its responses were partially grounded in evidence but were written as open-ended analyses rather than extractive answers, often with vague qualifiers such as "approximately" and "possibly," which reduced agreement with the gold answers. This performance degradation suggests that feeding long, dense Chinese financial passages into the context windows of these smaller multilingual models may overwhelm their instruction-following capabilities, leading to distracted generations where relying on their internal parametric knowledge occasionally yielded better heuristic guesses.

Cognitive Boundaries and Faithfulness Evaluation. Furthermore, the varied improvements across cognitive categories highlight the boundaries of standard RAG pipelines. While retrieval augmentation successfully surfaces explicit facts to drive statistically significant accuracy gains in Judgment Verification for both Qwen3-8B ($p=0.023$) and Qwen3-14B ($p=0.013$), improvements in tasks requiring multi-step reasoning or comparative calculation were universally non-significant across all six evaluated models. This demonstrates that merely supplying correct financial data does not endow the LLM with the necessary symbolic logic or arithmetic capabilities. Crucially, by decoupling gold-standard accuracy from our multi-dimensional evidence-support judgments, we prove that a large portion of "incorrect" answers are non-hallucinatory processing failures. Despite achieving low absolute accuracy scores, RAG models exhibited remarkably low true hallucination (fabrication) rates—such as 7.4% for Qwen3-14B and 8.2% for Qwen3-8B, confirming that multimodal RAG remains highly effective at constraining model confabulation in the financial domain.

8. Conclusion

Our study introduces CFQA, a benchmark for Chinese financial annual-report question answering, and presents baseline and retrieval-augmented results with detailed error analysis. Our evaluation demonstrates that RAG effectively improves answer accuracy for well-adapted models, most notably increasing Qwen3-14B's accuracy from 8.6% to 13.4%, and Qwen3-8B's accuracy from 9.6% to 12.0%. McNemar's tests reveal that the statistical significance of these improvements varies by model: Qwen3-14B achieved a highly significant overall gain ($p = 0.000126$), with notable task-specific significance in judgment verification ($p=0.013$) and enumeration ($p=0.043$). For Qwen3-8B, while the overall gain was not statistically significant ($p=0.082$), it still achieved statistically significant improvements specifically in the judgment verification task ($p=0.023$). Conversely, models like Mixtral-8x7B showed a statistically significant decrease in accuracy under RAG conditions ($p=0.012$). Furthermore, the system successfully controls hallucination rates, with RAG models exhibiting low fabrication rates ranging from 7.4% (Qwen3-14B) to 16.0% (Mixtral-8x7B). The proposed rule-based hallucination detector enables automated fact-checking and identifies evidence-contradictory claims with high precision, consistent with recent benchmarks (Bang et al., 2025; Sok et al., 2025). By explicitly addressing table and chart data alignment (Yang et al., 2025; Suri et al., 2025), this work helps address a gap in existing evaluations and offers a practical framework for developing more reliable domain-specific intelligent assistants.

Limitations

Despite the effectiveness of our RAG framework, several limitations warrant consideration. First, the evaluation is confined to financial annual reports and has a modest sample size ($N = 500$), which may limit generalization to open-domain contexts. Second, statistical analysis reveals that across all evaluated models, improvements in tasks requiring higher-order logic, such as "Comparative Calculation" and "Reasoning Analysis" consistently failed to reach statistical significance. This suggests that retrieval augmentation alone is insufficient to address LLMs' inherent deficits in multi-step logical reasoning and arithmetic operations. Furthermore, the system's approach to fusion remains preliminary, lacking deep semantic alignment between charts and text. The trade-off between retrieval precision and recall also requires optimization, as overly aggressive filtering may discard critical context.

Finally, our evaluation primarily relies on deterministic, rule-based metrics. While effective for verifiable factual consistency, this approach may not fully capture semantically correct answers with varied surface forms, potentially underestimating semantic faithfulness. Incorporating LLM-as-a-Judge frameworks could complement our current evaluation.

Ethics Statement

This research complies with ethical guidelines. All experimental data are from publicly available annual reports of listed companies and contain no Personally Identifiable Information (PII) or unauthorized trade secrets. Although our system improves financial fact-checking accuracy, given the probabilistic nature of Large Language Models, its outputs are for auxiliary reference only. They are not legally binding investment advice or audit conclusions.

9. Bibliographical References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. Benchmarking retrieval-augmented multimodal generation for document question answering. *arXiv preprint arXiv:2505.16470*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jeanie Genesis. 2025. Retrieval-augmented text generation: Methods, challenges, and applications. *Preprints*.

Ziyu Gong, Yihua Huang, and Chengcheng Mai. 2025. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering. *arXiv preprint arXiv:2508.00579*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*.

Channdeth Sok, David Luz, and Yacine Haddam. 2025. Metarag: Metamorphic testing for hallucination detection in rag systems. *arXiv preprint arXiv:2509.09360*.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of NAACL*, pages 6088–6109.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pages 5737–5762.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*.

Chaofa Yuan, Yuhan Liu, Rongdi Yin, Jun Zhang, Qinling Zhu, Ruibin Mao, and Ruifeng Xu. 2020. Target-based sentiment annotation in chinese financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5040–5045, Marseille, France. European Language Resources Association.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. In *Proceedings of EMNLP-IJCNLP*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

A. Design Characteristics of Each Question Type

A.1. Fact Extraction

- **Original (Chinese):** 2022 年中国人保的营业总收入是多少亿元?
- **Transliteration (Pinyin):** 2022 nián Zhōngguó Rénbǎo de yíngyè zǒng shōurù shì duōshǎo yì yuán?
- **Translation (English):** What was the total operating revenue of PICC in 2022, in hundreds of millions of yuan?

A.2. List Enumeration

- **Original (Chinese):** 列举 2022 年中国人保前五大客户的名称及其销售收入占比。
- **Transliteration (Pinyin):** Lièjǔ 2022 nián Zhōngguó Rénbǎo qián wǔ dà kèhù de míngchēng jí qí xiāoshòu shōurù zhànbǐ.
- **Translation (English):** List the names and sales revenue percentages of PICC's top five customers in 2022.

A.3. Comparison & Calculation

- **Original (Chinese):** 计算 2023 年中国神华营业收入的同比增长率和增长金额。
- **Transliteration (Pinyin):** Jìsuàn 2023 nián Zhōngguó Shénhuá yíngyè shōurù de tóngbǐ zēngzhǎng lǜ hé zēngzhǎng jīn'é.
- **Translation (English):** Calculate the year-on-year growth rate and growth amount of China Shenhua's operating revenue in 2023.

A.4. Judgment & Verification

- **Original (Chinese):** 2024 年建设银行是否存在商誉减值？如果是，减值金额是多少?
- **Transliteration (Pinyin):** 2024 nián Jiànshè Yínháng shìfǒu cúnzài shāngyù jiǎnzhí? Rúguǒ shì, jiǎnzhí jīn'é shì duōshǎo?
- **Translation (English):** Did CCB have goodwill impairment in 2024? If so, what is the impairment amount?

A.5. Reasoning & Analysis

- **Original (Chinese):** ROE 分解：分析 2024 年杜邦三因素对工商银行 ROE 变动的贡献。
- **Transliteration (Pinyin):** ROE fēnjiě: fēnxī 2024 nián Dùbāng sān yīnsù duì Gōngshāng Yínháng ROE biàndòng de gòngxiàn.
- **Translation (English):** ROE decomposition: Analyze the DuPont three-factor contribution to ICBC's ROE change in 2024.

B. Examples of the Hallucination Detector

Tables 4–6 provide illustrative examples of the hallucination detector's judgments, covering a correct answer, a potential hallucination, and a severe hallucination.

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的营业收入是多少? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de yíngyè shōurù shì duōshǎo? **Translation (English):** What was the operating revenue of Industrial and Commercial Bank of China (ICBC) in 2024?
Question Type	Fact Extraction
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。其中，利息净收入为 6789.12 亿元，手续费及佣金净收入为 1445.44 亿元。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. Qízhōng, lìxī jìng shōurù wéi 6789.12 yì yuán, shǒuxùfèi jí yōngjīn jìng shōurù wéi 1445.44 yì yuán. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%. Among this, net interest income was 6,789.12 billion yuan, and net fee and commission income was 1,445.44 billion yuan.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的营业收入为 8234.56 亿元，同比增长 3.2%。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de yíngyè shōurù wéi 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. **Translation (English):** The operating revenue of ICBC in 2024 was 8,234.56 billion yuan, representing a year-on-year growth of 3.2%.
Rule-Based Judgment	**Numerical Verification:** PASS (8,234.56 billion yuan found in evidence). **Text Coverage:** 0.85 (High coverage). **Reference Consistency:** PASS (Filename and page match). **Composite Score:** 0.92.
Final Verdict	**Verdict:** Evidenced. **Confidence:** 0.92. **Decision:** The answer is well-supported by evidence, no hallucination.

Table 4: Example 1: Correct Answer

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的净利润增长率是多少? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de jìnglìrùn zēngzhǎng lǜ shì duōshǎo? **Translation (English):** What was the net profit growth rate of ICBC in 2024?
Question Type	Fact Extraction
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。净利润为 3456.78 亿元。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. Jìnglìrùn wéi 3456.78 yì yuán. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%. The net profit was 3,456.78 billion yuan.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的净利润增长率为 5.8%，与上年相比有所改善。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de jìnglìrùn zēngzhǎng lǜ wéi 5.8%, yǔ shàngnián xiāngbǐ yǒu suǒ gǎishàn. **Translation (English):** The net profit growth rate of ICBC in 2024 was 5.8%, showing an improvement compared to the previous year.
Rule-Based Judgment	**Numerical Verification:** FAIL (5.8% not found in evidence). **Text Coverage:** 0.35 (Low coverage). **Reference Consistency:** PASS (Filename and page match). **Composite Score:** 0.48.
Final Verdict	**Verdict:** No Evidence. **Confidence:** 0.48. **Decision:** The answer contains information not found in the evidence, indicating hallucination risk.

Table 5: Example 2: Potential Hallucination

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的营业收入是同比增长还是下降? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de yíngyè shōurù shì tóngbǐ zēngzhǎng háishì xiàjiàng? **Translation (English):** Did ICBC's operating revenue increase or decrease year-over-year in 2024?
Question Type	Judgment Verification
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的营业收入同比下降了 2.5%，这主要是受市场环境的影响。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de yíngyè shōurù tóngbǐ xiàjiàng le 2.5%, zhè zhǔyào shì shòu shìchǎng huánjìng de yǐngxiǎng. **Translation (English):** ICBC's operating revenue decreased by 2.5% year-over-year in 2024, primarily due to market conditions.
Rule-Based Judgment	**Numerical Verification:** CONTRADICTION (Evidence indicates +3.2% growth, answer states -2.5% decline). **Text Coverage:** 0.25 (Very low coverage). **Reference Consistency:** PASS. **Composite Score:** 0.15.
Final Verdict	**Verdict:** Contradiction. **Confidence:** 0.15. **Decision:** The answer directly contradicts the evidence, indicating severe hallucination.

Table 6: Example 3: Severe Hallucination

Verifiable Financial Enterprise Question Answering via Inference-Time Grounding and Traceability

Anubha Kabra **Katie Jooyoung Kim** **Colin Zhiwei Kou**
Helene Sajer **Yimei Fan** **Gabriel Martinez Vidiri**
Bloomberg
{akabra16, jkim2425, ckou9, hsajer3, yfan258, gmartinezvi1} @bloomberg.net

Abstract

Financial enterprise AI systems deployed in high-stakes settings require responses that are verifiable, traceable, and auditable. We introduce a modular, model- and data-agnostic inference-time control framework, together with a deployment-aware evaluation strategy for verifiable financial enterprise question answering. Our method enforces faithfulness at inference time without retraining or changes to retrieval infrastructure. We deploy our method in a production financial enterprise assistant and evaluate it using a combination of intrinsic faithfulness metrics, baseline comparisons, and real-world user feedback. Our approach improves groundedness by 29% over baselines, reduces hallucinations to near-zero levels, and achieves near-perfect document-span traceability. Together, our results demonstrate that modular pipeline design combined with detailed, deployment-aware evaluation provides a practical and effective path toward verifiable financial enterprise QA systems.

Keywords: financial question answering, enterprise LLM systems, groundedness, span-level traceability, auditability, retrieval-augmented generation, faithful generation

1. Introduction

Institutions are increasingly deploying large language model (LLM) systems to support enterprise question answering across compliance, operations, product, and support functions (Huang et al., 2023; Gao et al., 2023). Prior work in financial NLP has explored domain-specific modeling and sentiment analysis for financial texts (Gao et al., 2023; Chen et al., 2024b; Huang et al., 2023), yet verifiable financial enterprise QA remains comparatively underexplored. In these high-stakes financial settings, responses must satisfy stricter standards than fluency or relevance alone; they must be verifiable, traceable to authoritative financial documents, and auditable under regulatory and operational scrutiny. Retrieval-augmented generation (RAG) pipelines, which pair LLMs with document retrievers to produce citation-backed responses, have shown strong performance on web-scale benchmarks (Kryscinski et al., 2020). However, despite their perceived interpretability, these systems often fall short of real-world verifiability and faithfulness requirements in financial enterprise QA. In regulated environments, users require not only fluent answers but guarantees that generated claims are factually supported, transparently verifiable, and robust to noisy or heterogeneous evidence sources (Chen et al., 2024b; Choubey et al., 2025).

Many failures in RAG systems arise at inference time rather than from retrieval or training deficiencies alone. During generation, LLMs must integrate evidence, synthesize claims, and assign citations under real-world constraints, yet existing pipelines offer limited support for auditing, monitoring, and intervention. As a result, deployed systems frequently produce unsupported claims, misattributed citations, or references that cannot be traced to concrete evidence spans, particularly when operating over long, unstructured enterprise documents (Joren et al., 2025; Choubey et al., 2025; Packowski et al., 2024). Such failures undermine interpretability, erode user trust, and hinder responsible deployment.

To study these challenges in practice, a month-long pilot deployment of an enterprise assistant built on a standard RAG architecture was conducted.

1.1. Pilot Deployment

The assistant used a state-of-the-art LLM[1] and retrieved the top-n documents from a heterogeneous corpus of internal sources, including wikis, policy manuals, knowledge base articles, and product documentation, using a customized retrieval backend optimized for low latency. Retrieved documents were incorporated into a task-specific prompt to generate natural language responses with inline citations.

The system was deployed for one month to 55 users across support, operations, compliance, and product roles, processing approximately 4,000 real-world queries. While users valued response fluency and relevance, the deployment revealed recurring failures such as incorrect citations, untraceable references, and unsupported claims.

Failure Analysis from Pilot Deployment

The pilot deployment revealed several recurring failure modes that limit the faithfulness of LLM-generated outputs in financial enterprise settings.

[1] Details withheld due to internal policies.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 39–48

Error Type	Description	Root Cause	Impact
Hallucinated Links	Nonexistent citations or URLs	Pattern-based generation	Erodes trust
Citation Drift	Cited passage doesn't support claim	Misaligned grounding	Reduces factual reliability
Limited Traceability	Hard to locate cited text	Buried content, weak anchors	Lowers transparency

Table 1: Key limitations of the pilot enterprise assistant deployment.

In particular, we observed the following caveats (See Table 1):

1. **Hallucinated links:** Generated citations or URLs that did not exist in the enterprise corpus.
2. **Citation drift:** Valid documents were cited but did not substantiate the associated claims.
3. **Limited traceability:** Even when citations were correct, verifying claims was difficult due to long or unstructured source documents.

To systematically analyze these failures, we characterize faithfulness along two dimensions. These are: (a) whether generated claims are factually supported by the cited source documents and (b) whether claims can be linked to specific, verifiable spans within those sources.

These observations motivate the following research questions:

RQ1: Can transparency into failure modes be incorporated into the pipeline design?

RQ2: Can we make LLMs more faithful by adding citation-level grounding and span-level traceability without retraining?

RQ3: Do improvements in LLM faithfulness translate to better downstream user satisfaction?

To address these research questions, we design and deploy `EvidenT`, a lightweight, post-hoc, model- and data-agnostic inference-time control system for extractive enterprise question answering. In light of **RQ1**, `EvidenT` adopts a modular pipeline design that enables step-wise inspection and evaluation of individual components, providing transparency into failure modes during system development. Addressing **RQ2**, the pipeline enforces citation-level groundedness and span-level traceability at inference time without requiring retraining or changes to retrieval infrastructure. This enforcement improves groundedness by 29%, achieves up to 99% span-level traceability, and reduces hallucinations to near-zero levels. With respect to **RQ3**, we introduce a detailed evaluation strategy that assesses faithfulness both prior to deployment and under real-world financial enterprise usage conditions, and observe that these targeted improvements are associated with improved downstream user satisfaction.

2. Related Work

2.1. Fine-tuning and Domain Adaptation

Fine-tuning LLMs has been widely explored to improve grounding in retrieval-augmented generation (RAG) systems (Huang et al., 2024; Penzkofer and Baumann, 2024; Zhang et al., 2024). However, such approaches require substantial domain-specific supervision, which is often impractical in financial enterprise settings. Moreover, fine-tuned models remain susceptible to hallucination or over-generalization under weak retrieval (Lee et al., 2025; Soudani et al., 2024), and primarily improve fluency rather than citation-level traceability (Ghosal et al., 2024). As a result, model-level optimization alone fails to guarantee transparent citation alignment or verifiable provenance (Ye et al., 2024; Huang et al., 2024), limiting applicability in high-stakes domains.

2.2. Basic RAG Systems

Despite their widespread use, RAG pipelines exhibit persistent failures in noisy and heterogeneous financial enterprise environments (Chen et al., 2024a). Retrieval remains a key bottleneck, as financial enterprise corpora are often fragmented, redundant, and inconsistently indexed (Sharma, 2025; Brown et al., 2025). Even when relevant documents are retrieved, *citation drift* – where cited passages do not support the generated claims – frequently occurs (Patel and Anand, 2024; Huang et al., 2024). Because retrieval and generation are loosely coupled, existing RAG systems lack explicit mechanisms to enforce span-level grounding, resulting in low citation precision and limited auditability.

2.3. LLM Faithfulness in real-world environments

While prior work has introduced citation-focused approaches and metrics to improve faithfulness in knowledge-grounded generation, for example in dialogue systems (Rashkin et al., 2021), more recent work has proposed citation-evaluation frameworks that assess citation quality in a more systematic way (Xu et al., 2025). However, much of this evaluation still works at a coarse level, often checking support only at the document or passage level. As a result, these approaches do not provide span-level source

attribution, which financial enterprise users need for fast and reliable verification. At the same time, widely used citation-generation benchmarks and setups (e.g., ALCE) are primarily grounded in public, research-style corpora and evaluation protocols (Gao et al., 2023), which can differ substantially from enterprise settings where collections are long, heterogeneous, and frequently unstructured (Anderson et al., 2024). Recent work on long-context settings further suggests that extracting and attributing evidence spans in long/unstructured inputs is itself challenging, reinforcing the need for span-level traceability beyond document-level citation correctness (Wright et al., 2025). As a result, current systems offer limited guarantees about where information comes from and do not fully support transparent, verifiable generation in real-world financial enterprise use.

3. Our Approach

Our approach consists of a model- and document-agnostic modular pipeline designed for robust citation generation in financial enterprise settings. An overview of how the pipeline processes queries is shown in Figure 1. The modular design aligns with RQ1, enabling transparent and incremental inspection of failure modes throughout pipeline development. This structure allows individual components to be analyzed and improved independently. The pipeline comprises the following components.

3.1. Document Retrieval Module

This module retrieves the top N relevant documents from heterogeneous financial enterprise data sources without relying on a centralized index. Instead, we retrieve documents independently from multiple distributed sources. Our solution is designed to be retrieval method-agnostic, enabling flexible integration with a variety of data sources.

3.2. Structured Passage Extraction Module

This module extracts candidate passages from retrieved documents using an LLM with structured output formatting. The prompt is designed to return verbatim spans from the document context. Each passage is represented in the following JSON structure:

```
{
  "passage_id": "<passage_id>",
  "url": "<url>",
  "content": "<passage_content>"
}
```

We utilize prior findings showing that LLMs achieve near-perfect performance on verbatim span extraction from long-context source documents (Hsieh et al., 2024) and structurally straightforward format transformations (Yang et al., 2026).

3.3. Source Alignment Filter Module

This module filters hallucinated passages based on n-gram overlap with their source documents. For each passage, we retrieve the full content of the source document using the provided `url` field. We then compute the n-gram overlap (typically $n = 5$) between the passage and the document content. Let p be an extracted passage and d its cited document. Define the n-gram overlap ratio as

$$\text{overlap}_n(p, d) = \frac{|G_n(p) \cap G_n(d)|}{|G_n(p)|}. \quad (1)$$

where $G_n(p)$ is the multiset of n-grams in p. We define the filtering action as follows:

- If **overlap** $>$ **threshold**, the passage is retained with the original `url`.
- If **0** $<$ **overlap** $\leq$ **threshold**, the passage is truncated to retain only the overlapping portion; the original `url` is preserved.
- If **overlap** $= 0$, we suspect citation drift. We iterate over all other retrieved documents to find one with overlap $>$ threshold. If found, the passage `url` is replaced with this.
- If none of these conditions apply, we drop the passage from the generated JSON.

To deliberately prioritize verifiability over textual abstraction, we employ an n-gram-based evaluation framework. We adopt lexical matching to accommodate the specialized finance domain, where documents are clause-driven and lexically precise. System identifiers, ticker symbols, and regulatory phrasing often carry specific operational meaning and cannot be freely paraphrased without altering intent (Kim et al., 2025; Li et al., 2025). These texts are structured and compliance-sensitive, differing substantially from open-domain text; the data is effectively out-of-distribution for semantic models. (Chen et al., 2024b; Anderson et al., 2024; Choi et al., 2025). Off-the-shelf semantic similarity models are trained on general-domain data and rely on subword tokenization, which can fragment rare financial identifiers and overlook clause-level distinctions that matter in compliance settings (Kudo and Richardson, 2018; Araci, 2019). Prior work shows that in such formulaic domains, lexical methods outperform semantic approaches (Choi et al., 2025; Thakur et al., 2021). In our early experiments, this approach showed superior performance compared to semantic metrics such as BARTScore, and also satisfied the strict latency requirements common in financial enterprise settings. In our setting,

the data distribution differs substantially from the training data of most semantic models. In practice, users can configure threshold values based on application requirements and desired alignment strictness. The framework remains extensible and can incorporate hybrid or semantic matching when domain-adapted semantic models are available.

3.4. Answer Generation Module:

A second LLM call generates the final answer using only the filtered and verified verbatim passages. Guided by a structured prompt, the LLM enforces strict grounding and formatting constraints to ensure maximal traceability while transforming the verbatim passages into a coherent, well-formed, and easily consumable answer without fragmentation. Each sentence in the final answer is cited using the associated URL fields, allowing the exact source location of the supporting content to be surfaced to the user. This design enhances user trust by enabling direct and transparent verification. In the user interface, we highlight the precise text spans corresponding to each citation in the source documents (see Figure 1). Clicking an inline citation takes the user directly to the exact location from which the referenced information originates.

3.5. Post-processing Module

The post-processing module programmatically refines the generation from the previous module to create a coherent final output to be presented to the users. The processing includes formatting paragraphs, removing repeated citation URLs, and normalizing company-specific terms to improve legibility.

3.6. Experimental Settings

We use two publicly available open-weight models of differing sizes – `LLaMA-3.1-8B` (M1) and `LLaMA-3.3-70B` (M2) – to demonstrate that our strategy is agnostic to model scale. We use open-weight models to adhere to data privacy constraints. The test set has approximately 500 financial enterprise QA data points, collected from our initial pilot study to closely reflect real user behavior. For maximal determinism, the temperature for each model call is set to 0.0.

4. Baselines

Building on the limitations in Section 2, we evaluate baselines that align with our goals of improving *faithfulness* without retraining or multi-stage orchestration.

Our design choices follow two principles. *(1) Scope alignment:* Our objective is to improve grounding and traceability in a model- and data-agnostic way; comparing with fine-tuned RAG systems (Asai et al., 2024; Lee et al., 2025) would conflate architectural complexity with our verification mechanism. *(2) Practical relevance:* Financial enterprise environments often preclude retraining or large-scale supervision due to privacy, fragmentation, and latency constraints, including strict limits on real-time LLM calls (Qian et al., 2025; Sun et al., 2024). Accordingly, we do not compare against verifier-based or self-reflective RAG systems, which rely on additional or iterative LLM calls and introduce latency, nondeterminism, and auditability challenges that are incompatible with our financial enterprise deployment constraints. We evaluate under realistic plug-and-play conditions. `EvidenT` remains complementary to advanced verifier-based systems and can be layered atop them for stronger grounding and traceability. We compare against two representative baselines:

Direct Prompting with Inline Citations: The LLM generates citations inline as part of its response, serving as a minimal citation-aware baseline without explicit verification (Singal et al., 2024; Lewis et al., 2020).

Ground-Every-Sentence: The LLM appends a citation after every sentence, ensuring each atomic statement is grounded in at least one retrieved document, enabling fine-grained evaluation of citation precision and coverage (Xia et al., 2025).

5. Evaluation

While the ultimate measure of an enterprise assistant's success lies in **user satisfaction**, such outcomes can only be reliably assessed post-deployment. During development, we therefore rely on *proxy metrics* that correlate with user trust and perceived reliability. Below, we present a detailed evaluation setup and results comparing our proposed approach with the two baselines introduced in Section 4. We report: (a) an *intrinsic evaluation*, (b) *comparative results* against baselines; and (c) *post-deployment metrics* based on real-world user feedback. Together, *(a)* and *(b)* form a practical proxy evaluation strategy for assessing groundedness, traceability, and trustworthiness during development, which can be applied to other modular pipelines like ours.

5.1. Intrinsic Evaluation (RQ1)

The pipeline is designed to support incremental evaluation based on observed failure modes, enabling targeted analysis of individual components rather than treating the system as a whole. This design allows us to isolate and address problematic

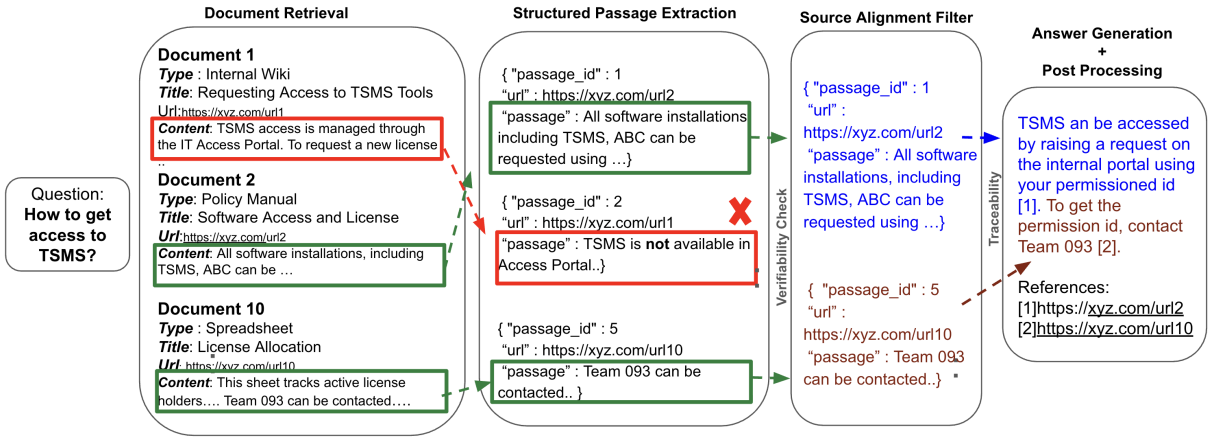


Figure 1: A step-by-step visual of our pipeline showing how a query is processed, from *document retrieval* to *answer generation* with example content that is entirely fictional and used only for illustration.

steps more effectively.

5.1.1. Evaluating Source Alignment Filter Module

The goal is to assess whether the passages generated in this module are factually aligned with the retrieved reference documents or not. To this end, we quantify two specific error categories: `URL` *drift* and *hallucinations*.

- **% citation drift:** The percentage of cases where the generated `URL` does not match the reference `URL`, yet corresponds to a valid `URL` present in the retrieved documents.
- **% hallucination:** The percentage of generated passages that contain one or more hallucinations.

As shown in Table 2, we observe a considerable amount of citation drift. However, we are able to locate the correct citation within the retrieved documents that matches the passage and substitute accordingly. The larger model demonstrates stronger factual grounding, exhibiting significantly fewer hallucinated passages. Notably, when hallucinations do occur, the initial portion of the generated response is often accurate, but as the generation continues, it gradually diverges from the source material.

5.1.2. Human Evaluation of Generated Passages

After generating the source passages (*Structured Passage Extraction Module*) and processing them through the filtering (*Source Alignment Filter*), we conducted human evaluation to assess the relevance of the passages to the original queries. Evaluating relevance required *subject-matter expertise*, a capability that current LLMs do not inherently possess.

	M1	M2
% citation drift	24.19	13.5
% hallucinated	16.4	0.6

Table 2: N-gram filter metrics for models M1 and M2.

	M1	M2
% Relevant	42	59
% Partially relevant	58	87

Table 3: Human evaluation results of passage relevance for models M1 and M2.

Table 3 presents, for each query, the number of passages that evaluators judged as fully or partially relevant, which we then averaged across all queries. Two annotators reviewed 50 queries and corresponding passages, and we compute the final score by averaging their judgments.

The smaller model M1 returns a higher share of irrelevant passages, while the larger model M2 consistently produces more fully or partially relevant results, achieving up to 87% partially relevant passages. Both models can surface relevant passages from noisy, unstructured documents, indicating the pipeline effectively identifies meaningful content despite input noise. The second LLM call in the answer generation module further chooses what to surface from these passages, adding another layer of verification to the source documents.

5.2. Evaluating Faithfulness (RQ2)

5.2.1. Evaluating Groundedness

To evaluate citation groundedness in relation to RQ2, we compute the following metrics:

%Hallucination: A binary measure indicating whether an answer contains any hallucinated citations. In financial enterprise settings like ours, even

Method	%Groundedness		%Hallucination	
	M1	M2	M1	M2
Direct Prompting	52	73	31	18
Ground Every Sentence	47	71	54	23
EvidenT	**77**	**93**	**0**	**0**

Table 4: Comparison across groundedness and hallucination metrics.

Method	SemMax		SemRecall	
	M1	M2	M1	M2
Direct Prompting	3.280	4.100	-0.390	1.930
Ground Every Sentence	4.168	4.320	0.439	2.360
EvidenT	**5.360**	**5.990**	**4.390**	**5.450**

Table 5: Semantic similarity metrics across different approaches and model variants.

a single hallucinated citation is unacceptable; therefore, an answer receives a score of 0 if it contains any citation that does not correspond to a retrieved document, and a score of 1 otherwise. Let c' denote the set of citations generated in an answer and c denote the set of all retrieved URLs. Then:

$$\text{Hallucination} = \mathbb{1}\big[\,\mathrm{c}' \subseteq \mathrm{c}\,\big] \tag{2}$$

where $\mathbb{1}[\cdot]$ is equal to 1 if all generated citations exist within the retrieved documents, and 0 if any are hallucinated.

%Groundedness: A measure of whether the model's answer is substantively supported by authoritative source documents without introducing unsupported citations. We use human-annotated URLs that are sufficient to answer each test instance. Because financial enterprise QA operates in an open-world setting where multiple documents may independently support the same answer, the gold set is sufficient but not exhaustive.

An answer is considered grounded if (i) at least one generated citation overlaps with the expert-annotated sufficient set, and (ii) no generated citation is hallucinated (i.e., all citations correspond to retrieved documents). This definition preserves the open-world assumption while enforcing citation validity, yielding a stricter and deployment-aligned measure of faithfulness.

Let c' denote the set of citations generated in the model's output and c denote the set of gold citations. We define groundedness as:

$$\text{Groundedness} = \mathbb{1}\big[\,\mathrm{c}' \cap \mathrm{c} \neq \emptyset\,\big] \tag{3}$$

where $\mathbb{1}[\cdot]$ is 1 if there is any overlap between the gold-labelled and generated citations, and 0 otherwise.

As shown in Table 4, the EvidenT approach yields a substantial improvement in factual groundedness while completely eliminating hallucinations. This improvement can be largely attributed to the *N-gram Filtering Module* (*Source Alignment Filter*), which proactively filters out hallucinated or citation-drift passages prior to generating the final response. In contrast, the baseline prompting strategies exhibit a notable degree of both hallucination and citation drift. Between the two model variants, M2 consistently outperforms M1 in both groundedness and hallucination resistance. Overall, EvidenT achieves a **29%** relative improvement in groundedness for M2, with zero instances of hallucination observed across the evaluated set.

5.2.2. Evaluating Traceability

In addition to the above, we use both semantic and lexical post-generation metrics to quantify how well an answer generated via EvidenT can be traced back to its source document at the span level pertaining to RQ2.

For all subsequent evaluations, we first extract factual statements by taking the text preceding each citation: for example, the blue and brown spans in Figure 1 illustrate two separate facts. For EvidenT, we additionally gather both the extracted facts and their corresponding source passages by matching cited URLs. These are then compared against the content of the cited documents, identified through the same cited URLs. Results are averaged across all queries.

Given a reference document D and a model-generated answer A, we quantify how traceable A is to D. We employ several methods to measure this. Let U and V denote the multisets of tokens from A and D, respectively. t denotes the token.

5.2.3. Word Overlap

Let $O = \sum_t \min(\text{count}_U(t), \text{count}_V(t))$ for each token t.

$$\text{AnsCov} = \frac{O}{|U|}, \qquad \text{DocFocus} = \frac{O}{|V|} \tag{4}$$

AnsCov measures how well the document covers the answer content, while DocFocus reflects how concentrated the document is on that answer.

5.2.4. n-gram Overlap

For $n \in \{2, 3, 5, 10\}$, let $\mathcal{G}_n(T)$ denote the multiset of n-grams in text T, and let $I_n = |\mathcal{G}_n(A) \cap \mathcal{G}_n(D)|$ be the number of overlapping n-grams between the answer A and document D. We define AnsCov@n, which measures how much of the answer's phrasing is covered by the document, and DocFocus@n, which captures how concentrated the document is on the answer content.

Method	Model	DocFocus	AnsCov	Focus@2	Cov@2	Focus@3	Cov@3	Focus@5	Cov@5	Focus@10	Cov@10
Direct Prompting	M1	0.053	0.733	0.031	0.457	0.025	0.377	0.021	0.331	0.014	0.292
	M2	0.047	0.887	0.032	0.644	0.026	0.528	0.020	0.436	0.012	0.324
Ground Every Sentence	M1	0.044	0.736	0.027	0.442	0.022	0.363	0.019	0.309	0.014	0.265
	M2	0.039	0.897	0.028	0.655	0.023	0.536	0.018	0.431	0.011	0.328
EvidenT	M1	**0.096**	**0.998**	**0.093**	**0.996**	**0.093**	**0.995**	**0.091**	**0.993**	**0.096**	**0.986**
	M2	**0.162**	**0.999**	**0.162**	**0.996**	**0.160**	**0.993**	**0.158**	**0.990**	**0.156**	**0.979**

Table 6: Answer Coverage (*AnsCov*) and Document Focus (*DocFocus*) metrics across methods and models. Higher *Coverage* indicates that a document captures more of the generated answer's content, while higher *Focus* reflects a greater proportion of the document being relevant to the answer.

$$\text{AnsCov@}n = \frac{I_n}{|\mathcal{G}_n(A)|}, \quad \text{DocFocus@}n = \frac{I_n}{|\mathcal{G}_n(D)|} \tag{5}$$

Larger n values emphasize near-verbatim phrasing and reduce tolerance for paraphrasing. Results are shown in Table 6. The high AnsCov values and low DocFocus values reflect the long and noisy nature of the source documents. EvidenT substantially improves coverage over both baselines. In particular, at $n = 10$, EvidenT achieves near-saturated scores (AnsCov: 0.999 vs. 0.897; Cov@10: 0.979 vs. 0.328), indicating that generated answers preserve long contiguous spans from the cited documents.

In compliance-sensitive financial settings, this degree of span-level alignment is operationally important: users must verify claims directly against authoritative clauses. The strong lexical coverage therefore reflects audit-grade traceability rather than incidental surface overlap.

5.2.5. Semantic Matching

For semantic scoring, we create overlapping windows (256 tokens, stride 50) for both A and D. We use the off-the-shelf sentence cross-encoder to compute similarity scores between each answer $\{a_i\}$ and each document window $\{d_j\}$. The model jointly encodes each text pair and outputs a scalar relevance score, which we use without normalization.

$$\text{SemMax} = \max_{i,j} s(a_i, d_j), \tag{6}$$

$$\text{SemRecall} = \frac{1}{|\{a_i\}|} \sum_i \max_j s(a_i, d_j) \tag{7}$$

where $s(a_i, d_j)$ denotes the raw relevance score produced by the cross-encoder. SemMax captures the strongest localized semantic match between the answer and the document, while SemRecall reflects the overall semantic coverage of A by D.

Table 5 shows that SemMax scores are consistently higher than SemRecall across all methods and model sizes. This indicates that generated answers generally contain at least some segments that align well semantically with the source documents. However, when semantic alignment is averaged across all answer segments (i.e., SemRecall), weaker methods, particularly for M1, exhibit substantially lower coverage, suggesting fragmented or inconsistent grounding. Across both model sizes, EvidenT achieves the strongest performance on both metrics. For M2, EvidenT improves SemMax by approximately 39% relative to the strongest baseline and yields more than a 100% improvement in SemRecall. Similar trends hold for M1, where gains are even more pronounced for SemRecall. These results indicate that EvidenT not only produces answers with stronger localized semantic matches, but also maintains significantly more consistent semantic alignment with the source documents overall.

5.2.6. LLM-as-a-Judge Entailment

We employ an LLM-as-a-judge by explicitly reformulating the evaluation as a fine-grained entailment task. Instead of relying on a holistic or impressionistic judgment over an entire answer, we decompose the output into individual factual statements in A. For each statement, the model assigns a ternary label of *Yes*, *Partial*, or *No*, depending on whether the cited source document fully supports, partially supports, or does not support the claim. This formulation sharply constrains the role of the LLM and eliminates much of the ambiguity inherent in broad LLM-as-a-judge setups. In this setting, the LLM effectively acts as a semantic entailment model grounded in explicit evidence, rather than as an unconstrained evaluator. We use a different off-the-shelf LLM for this assessment and report results using only the larger model (M2), given its clear advantage over M1. As shown in Table 7, EvidenT achieves the highest entailment accuracy, with substantially more fully supported (Yes) statements and the fewest unsupported (No) ones, demonstrating stronger factual alignment with cited sources than all baselines.

5.3. Post-Deployment Evaluation (RQ3)

EvidenT explicitly targets improvements in faithfulness; we therefore evaluate whether gains along this dimension are associated with changes in

Method	% Yes	% Partial	% No
Direct Prompting	59.41	**27.58**	8.22
Ground Every Sentence	56.35	27.25	11.27
EvidenT	**74.38**	23.14	**1.65**

Table 7: Entailment metrics from LLM-as-a-Judge.

Metric	Pilot	EvidenT
Helpfulness	64%	92%
Relevance	75%	89%

Table 8: Post-deployment user metrics.

user satisfaction following deployment. We analyze user feedback collected during live usage of the financial enterprise assistant before and after introducing EvidenT, focusing on two metrics: *helpfulness*, measured via mandatory thumbs-up / thumbs-down feedback, and *relevance*, collected as a follow-up signal for helpful responses.

As shown in Table 8, deployment of EvidenT is followed by substantial improvements across both metrics. Helpfulness increases by 28%, while relevance improves by 19%. In addition, we observe a post-deployment user retention rate of 90%, indicating sustained engagement with the system.

While these metrics reflect overall user experience and may be influenced by multiple system-level factors, the targeted nature of the pipeline changes suggests that improvements in faithfulness contribute meaningfully to the observed gains. We do not claim causality; rather, the results indicate a strong association between increased faithfulness and improved user satisfaction in a real-world financial enterprise setting.

6. Conclusion

We presented a modular inference-time pipeline designed to improve faithfulness in financial enterprise extractive question answering systems. Beyond improving verifiability, the modular structure of the pipeline enables fine-grained inspection and debugging of individual components, facilitating the identification and mitigation of failure modes during system development and deployment (**RQ1**). In parallel, we introduced a comprehensive evaluation strategy that combines intrinsic metrics, comparative baselines, and post-deployment user feedback to assess groundedness, traceability, and downstream impact under real-world constraints. Our results demonstrate that citation-level groundedness and span-level traceability can substantially improve factual faithfulness without requiring model retraining (**RQ2**), and that these improvements translate into measurable gains in downstream user satisfaction (**RQ3**). Together, our findings demonstrate that modular pipeline design, coupled with detailed and stage-aware evaluation, provides a practical and effective strategy for building trustworthy LLM systems in industry settings.

We emphasize that the contribution of this work lies not in proposing algorithmic novelty, but in demonstrating that carefully designed inference-time control mechanisms can deliver verifiable, auditable behavior under strict latency and operational constraints. In high-risk financial enterprise environments, where retraining, multi-stage verification loops, and nondeterministic generation may be impractical, such engineering-oriented interventions provide a scalable and deployment-aligned path toward trustworthy LLM systems.

7. Ethical Considerations and Limitations

Our study is limited to text-based financial enterprise extractive question answering and focuses on groundedness and traceability as the primary optimization objectives. Accordingly, we do not evaluate free-form reasoning, creative generation, or standard public benchmarks, as these settings do not reflect financial enterprise operational and trust constraints. Due to data privacy requirements, the underlying financial enterprise datasets cannot be publicly released. We rely on lexical and semantic traceability metrics as proxies for downstream trust; controlled causal analyses, such as A/B testing, are out of scope for this work, although post-deployment user feedback provides an initial signal of real-world impact. Additionally, our experiments are restricted to open-weight models due to deployment and privacy constraints, and we do not consider multimodal inputs or non-textual evidence. Extending the pipeline and evaluation framework to these settings remains future work.

8. Bibliographical References

Eric Anderson, Jonathan Fritz, Austin Lee, Bohou Li, Mark Lindblad, Henry Lindeman, Alex Meyer, Parth Parmar, Tanvi Ranade, Mehul A Shah, et al. 2024. The design of an LLM-powered unstructured analytics system. *arXiv preprint arXiv:2409.00847*.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint arXiv:2508.06401*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhiyu Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Ruth Petzold, and William Yang Wang. 2024b. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*. Survey Certification.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. 2025. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the ICLR 2025 Workshop on Advances in Financial AI*.

Prafulla Kumar Choubey, Xiangyu Peng, Shilpa Bhagavath, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. 2025. Benchmarking deep search over heterogeneous enterprise data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Industry Track)*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. Understanding finetuning for factual knowledge extraction. In *International Conference on Learning Representations*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. 2025. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems. *arXiv preprint arXiv:2503.15191*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, pages 66–71.

Zhan Peng Lee, Andre Lin, and Calvin Tan. 2025. Finetune-rag: Fine-tuning language models to resist hallucination in retrieval-augmented generation. *arXiv preprint arXiv:2505.10792*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Xiaochen Li, Domenico Bianculli, and Lionel Brian. 2025. Tracing content requirements in financial documents using multi-granularity text analysis. *Requirements Engineering*, 30(1):109 – 132.

Sarah Packowski, Inge Halilovic, Jenifer Schlotfeldt, and Trish Smith. 2024. Optimizing and evaluating enterprise retrieval-augmented generation (rag): A content design perspective. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pages 162–167.

Maya Patel and Aditi Anand. 2024. Factuality or fiction? benchmarking modern llms on ambiguous qa with citations. *arXiv preprint arXiv:2412.18051*.

Vinzent Penzkofer and Timo Baumann. 2024. Evaluating and fine-tuning retrieval-augmented language models to generate text with accurate citations. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. 2025. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *SIGIR-AP*, pages 47–54.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Chaitanya Sharma. 2025. Retrieval-Augmented Generation: A comprehensive survey of architectures, enhancements, and robustness. *arXiv preprint arXiv:2506.00054*.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs. In *Proceedings of the Fact Extraction and VERification Workshop (FEVER)*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine-tuning vs. retrieval-augmented generation for less popular knowledge. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (Asia-Pacific)*.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the NeurIPS 2021 Track on Datasets and Benchmarks*.

Dustin Wright, Zain Muhammad Mujahid, Lu Wang, Isabelle Augenstein, and David Jurgens. 2025. Unstructured evidence attribution for long context query focused summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025. Ground Every Sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics*.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. 2025. Citeeval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32759–32778.

Jialin Yang, Dongfu Jiang, Tony He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. 2026. StructEval: Benchmarking LLMs' capabilities to generate structural outputs. *Transactions on Machine Learning Research*.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language models to domain-specific rag. In *Conference on Language Modeling*.

Environmental, Social and Governance Sentiment Analysis on Slovene News: A Novel Dataset and Models

Paula Dodig[1] **Boshko Koloski**[2] **Katarina Sitar Šuštar**[3]
Senja Pollak[2] **Matthew Purver**[2,4]
[1] Eindhoven University of Technology, Eindhoven
[2] Jožef Stefan Institute and Postgraduate School, Ljubljana
[3] Faculty of Economics, University of Ljubljana
[2] Queen Mary University of London
p.dodig@student.tue.nl, {boshko.koloski, senja.pollak}@ijs.si,
katarina.sitar@ef.uni-lj.si, m.purver@qmul.ac.uk

Abstract

Environmental, Social, and Governance (ESG) considerations are increasingly integral to assessing corporate performance, reputation, and long-term sustainability. Yet, reliable ESG ratings remain limited for smaller companies and emerging markets. We introduce the first publicly available Slovene ESG sentiment dataset and a suite of models for automatic ESG sentiment detection. The dataset, derived from the MaCoCu Slovene news collection, combines large language model (LLM)-assisted filtering with human annotation of company-related ESG content. We evaluate the performance of monolingual (SloBERTa) and multilingual (XLM-R) models, embedding-based classifiers (TabPFN), hierarchical ensemble architectures, and large language models. Results show that LLMs achieve the strongest performance on Environmental (Gemma3-27B, F1-macro: 0.61) and Social aspects (gpt-oss 20B, F1-macro: 0.45), while fine-tuned SloBERTa is the best model on Governance classification (F1-macro: 0.54). We then show in a small case study how the best-preforming classifier (gpt-oss) can be applied to investigate ESG aspects for selected companies across a long time frame.

Keywords: sentiment analysis, ESG, economics, environment, social, governance, large language models, dataset, single-task, multi-task, transformers, financial NLP

1. Introduction

Environmental, Social, and Governance (ESG) considerations have become essential in the evaluation of corporate performance and investment potential (Chen et al., 2023). Increased awareness of corporate sustainability has led to the integration of ESG metrics into financial and public evaluations of businesses. Despite this momentum, a significant number of smaller publicly listed companies lack formal ESG ratings, making it difficult for ESG-focused retail investors to assess their sustainability performance (Bazrafshan, 2023). Moreover, existing ESG ratings are typically static, updated infrequently, and therefore unable to capture short-term shifts in public or media perception. They also tend to aggregate information from limited, often homogeneous sources, which obscures variation in how companies are portrayed across different news outlets and domains. Consequently, traditional ESG ratings fail to provide a dynamic or diversified view of corporate reputation as it evolves in real time. This gap is particularly pronounced in less-resourced linguistic contexts, where limited data availability and language-specific barriers further hinder analysis. Our research addresses this challenge by developing an automated, sentiment-based framework leveraging large language models (LLMs) to evaluate ESG-related content in news articles, with a specific focus on Slovene — a less-resourced Slavic language.

The research addresses the following questions. First, can we develop an automated, sentiment-based framework for ESG aspects in Slovenian textual data, more specifically in Slovenian news? Second, can we use this framework to track ESG-related perception of companies through associated news text?

The main contributions of this paper are as follows. First, we present the first sentiment-annotated dataset from Slovenian news media on the aspects of Environment, Social and Governance considerations (**SloESG-News 1.0**). The development of this dataset, is based on the MaCoCu Slovene News dataset (Bañón et al., 2022) and uses the IPTC news codes and large language models (LLMs) for selection of articles for annotation. The gold standard annotation is provided by human annotators, resulting in a new publicly available resource for Slovenian. Next, the dataset is used for training **ESG sentiment models** for Slovenian and evaluating their performance. By extensive set of experiments using fine-tuned monolingual and multilingual transformers, LLMs, embedding-based classifier and hierarchical ensembles, we provide a replicable methodology and select the best models for each aspect. Third, we use the selected models in a **case study** with an expert from the field of economics. We apply the ESG model on a corpus of Slovene news for selected companies

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 49–58

across 15 years, and show the change in E, S and G sentiment across time. The contextualisation and interpretation of results shows the potential of our method for interdisciplinary research and further more detailed qualitative case studies.

2. Related Work

The increasing relevance of ESG topics has driven the development of computational methods for understanding sustainability discourse in text. Prior research on ESG-related text analysis has focused on company reports and financial disclosures, leveraging supervised machine learning to assess sentiment and topic relevance (Nassirtoussi et al., 2015). However, such approaches are often limited by the availability of labeled data and by their focus on English and other high-resource languages.

Recent advances in transformer-based language models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have enabled more nuanced sentiment and topic classification across domains, including finance and social responsibility (Araci, 2019).

The application of NLP to ESG analysis has gained significant momentum, with transformer-based models proving particularly effective for processing corporate disclosures and news articles at scale (Schimanski et al., 2024). Domain-specific models like FinBERT-ESG and ESGBERT have been developed through fine-tuning on ESG-specific corpora, achieving strong performance across environmental, social, and governance classification tasks (Araci, 2019; Mehra et al., 2022). Recent work has explored knowledge-enhanced approaches, with Koloski et al. (2022) proposing representations that combine knowledge graphs and taxonomies with document embeddings for sustainability detection, while Angioni et al. (2024) employed knowledge graphs to track ESG discourse evolution in news articles. BERT-based sentiment analysis has demonstrated predictive power for market reactions, with positive ESG news correlating with average abnormal returns of 0.31% and negative news with -0.75% (Dorfleitner and Zhang, 2024). The FinNLP workshop series has hosted multilingual ESG research through shared tasks on ESG issue identification across multiple languages (Tseng et al., 2023), while recent studies have integrated ESG sentiment with technical indicators for financial forecasting (Lee et al., 2024). However, most work focuses on English and high-resource languages, with limited research on low-resource contexts like Slovene.

Recenlty, LLMs have been explored as tools for dataset curation and pseudo-labeling. The teacher–student framework for topic classification (Kuzman and Ljubešić, 2025) has been shown to produce reliable results with minimal manual supervision, especially for under-resourced languages. Building on these insights, our study applies LLM-assisted filtering and human validation to create the first Slovene ESG sentiment dataset.

3. SloESG-News 1.0 dataset

To create an appropriate dataset, we extracted articles from the MaCoCu Slovenian dataset (Bañón et al., 2022), filtering for a curated list of Slovenian companies, defined by an expert in ESG focusing on companies where at least one of the three aspects E, S or G is strongly present. The data was preprocessed to extract a subset of news articles where company names and ESG-related terminology co-occurred, by applying the Slavic-XLMR named entity recognition model (Ivačič et al., 2023), together with the IPTC media topic classifier from the CLASSLA repository (Kuzman and Ljubešić, 2025). Manual annotation was then conducted in collaboration with economics students from the University of Ljubljana, who were trained to tag sentiment (positive, neutral, negative, or irrelevant) separately for each ESG aspect (Environmental, Social, Governance). The sentiment label was assessed specifically from the point of view of the company mentioned in the text (thus necessitating the inclusion of the "irrelevant" category). Initially, 24 annotators were considered, each given a set of 40 articles, 30 unique for individual annotation, and 10 shared articles jointly annotated by everyone. Along with student annotators, an expert annotation was used to identify outliers. After a student-expert pairwise agreement was calculated, 6 of the student annotators were discarded from the dataset due to a low agreement level, signifying faulty annotations and outlier behavior. The final dataset consists of 550 unique articles, where 10 articles were annotated by 19 annotators, while 540 by a single annotator.

Ten articles were annotated by all annotators, allowing us to calculate inter-annotator agreement using Fleiss' *kappa* metric for multiple annotators (Fleiss and Cohen, 1973). This highlighted the inherent complexity of ESG sentiment interpretation: while the E category showed very strong agreement (close to 0.8), S agreement was in the moderate range (0.4) and we saw only low agreement on the G category (0.2). A heatmap of sentiment counts can be seen in Gigure 1.

The dataset is split into training and test parts (see Table 1) and will be made available on CLARIN upon acceptance.

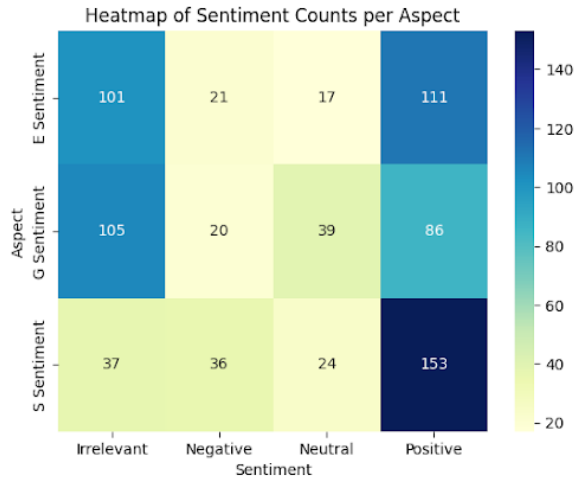


Figure 1: Student annotation results

Table 1: Dataset distribution.

Split	Aspect	Irrel.	Neg.	Neut.	Pos.
Train (440)	E	288	41	39	72
	S	144	48	69	179
	G	193	78	86	83
Test (110)	E	77	6	12	15
	S	37	15	22	36
	G	53	23	19	15

4. Methodology for ESG modelling

Our methods used to classify ESG-related sentiment on the proposed dataset focus on two different perspectives: adapting pre-trained machine learning models (such as BERT and TabPFN) and zero-shot querying of LLMs, ranging from the monolingual Slovene model GaMS to the multilingual reasoning model GPT-OSS, as well as building a hirearchically stacked ESG model.

Our approach follows a multi-level stacking paradigm consisting of three principal stages: a) feature extraction through multiple text representation methods, b) base-level classification using diverse model families, and c) meta-level prediction through hierarchical neural ensembles. The complete pipeline is illustrated in Figure 2.

4.1. Text Representation Models

We employ five distinct text encoding strategies.

4.1.1. Multilingual Sentence Embeddings

BGE-M3[1] (BAAI/bge-m3): A state-of-the-art multilingual embedding model supporting over 100 languages. The model produces dense 1024-dimensional vectors optimized for semantic similarity tasks through contrastive learning on large-scale multilingual corpora.

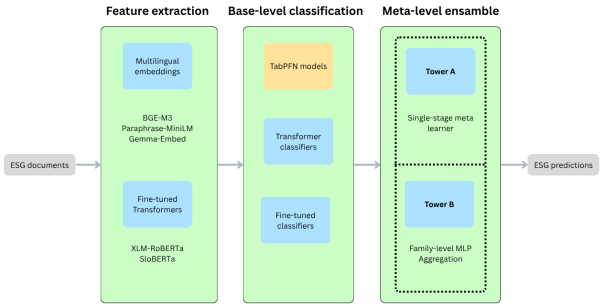


Figure 2: Methodology pipeline

Paraphrase-Multilingual-MiniLM-L12-v2[2]: A distilled sentence transformer architecture based on the MiniLM framework (Wang et al., 2020), providing computationally efficient 384-dimensional embeddings.

Gemma-Embed[3] (google/embeddinggemma-300m): A task-agnostic embedding model built on Google's Gemma architecture (Gemma Team, 2025).

4.1.2. Fine-tuned Transformer Classifiers

XLM-RoBERTa-base (Conneau et al., 2020): A cross-lingual pre-trained transformer model trained on 2.5TB of CommonCrawl data covering 100 languages. Unlike multilingual BERT, XLM-RoBERTa employs no language-specific embeddings, instead learning cross-lingual representations through language-agnostic pretraining. We fine-tune all layers on the ESG classification task with a linear classification head (768 $\rightarrow$ 4 classes per aspect).

SloBERTa (Ulčar and Robnik-Šikonja, 2021): A RoBERTa variant specifically pre-trained on Slovenian texts. This model provides specialized morphological and syntactic knowledge for Slovenian, a highly inflected South Slavic language. The architecture mirrors RoBERTa-base with language-specific tokenization and vocabulary.

4.1.3. Dimensionality Reduction

To address TabPFN's computational constraints on high-dimensional inputs, we optionally apply Truncated Singular Value Decomposition (SVD) to the embedding matrices. We evaluate candidate dimensions $\mathcal{D} = \{32, 64, 128, 256\}$ through nested validation, selecting the dimensionality $d^* \in \mathcal{D}$ that

[1] https://huggingface.co/BAAI/bge-m3

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[3] https://huggingface.co/google/embeddinggemma-300m

Hyperparameter	Value
Optimizer	AdamW
Learning rate	3×10^{-5}
Weight decay	0.01
Batch size	128
Max sequence length	192 tokens
Warmup steps	100
Learning rate schedule	Linear
Max epochs	100
Early stopping patience	15 epochs
Metric for model selection	Macro-F1

Table 2: Hyperparameters for transformer fine-tuning.

maximizes macro-averaged F1 score on a held-out 20% internal validation split from the training partition.

4.2. Base-Level Classification

The assign positive, negative, neutral or irrelevant label for each ESG category.

4.2.1. TabPFN-based Models

TabPFN (Prior-Fitted Networks) (Hollmann et al., 2022) is a meta-learned classifier that performs approximate Bayesian inference through in-context learning without gradient-based training. Given embedding matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and labels $\mathbf{y}$, TabPFN produces probabilistic predictions $p(\mathbf{y}^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$ for test instances $\mathbf{x}^*$ through a single forward pass, leveraging patterns learned from synthetic tabular datasets during meta-training.

We construct six TabPFN-based classifiers by pairing three embedding models (BGE-M3, Paraphrase-MiniLM, Gemma-Embed) with two preprocessing variants (with/without SVD). Each classifier operates independently on the E, S, and G aspects, producing 4-class probability distributions. Model identifiers follow the convention `tabpfn_{embedding}` and `tabpfn_{embedding}_svd`.

4.2.2. Fine-tuned Transformer Models

Transformer models are trained as sequence classification systems using the Hugging Face Trainer API (Wolf et al., 2020) with the hyper-parameters specified in Table 2. Each aspect ($a \in \{\mathrm{E}, \mathrm{S}, \mathrm{G}\}$) is trained independently as a 4-class classification task (pos, neg, neut, irr), resulting in aspect-specific fine-tuned models. Model identifiers follow the convention `hf_(SloBERTa/XLMR)`.

4.3. Large Language Models

To assess the performance of instruction-following LLMs in zero-shot and few-shot ESG sentiment classification, we evaluate five models from the Gemma and GPT families. Unlike fine-tuned transformers, these models are prompted to classify ESG sentiment without gradient-based adaptation. We employ a structured prompt template that presents the classification task with explicit ESG definitions and class descriptions. We evaluate models of varying parameter counts to assess the impact of scale on ESG classification:

- **GaMS-9B / GaMS-27B** (Vreš et al., 2024): Gemma-based models fine-tuned for Slovenian language understanding
- **Gemma3-12B / Gemma3-27B** (Gemma Team, 2025): Instruction-tuned variants from the Gemma 3 family
- **gpt-oss 20B** (OpenAI Team, 2025): An open-source GPT-architecture reasoning model with 20B parameters

Inference Configuration: For certain models, we explore few-shot prompting by including $k \in \{10, 20\}$ labeled examples in the prompt context (denoted by model suffix, e.g., Gemma3-12B). Temperature is set to 0.0 for deterministic outputs, and responses are parsed to extract class predictions for each ESG aspect.

LLMs are evaluated directly on the test set $\mathcal{D}_{\text{test}}$ without additional training, providing a comparison baseline for zero-shot transfer performance against fine-tuned and ensemble approaches.

4.4. Meta-Feature Construction

Base model predictions are transformed into meta-features through the following pipeline:

1. **Probability Extraction**: Each base model produces probability distributions $\mathbf{P}_a \in \mathbb{R}^{n \times 4}$ for aspect $a \in \{\mathrm{E}, \mathrm{S}, \mathrm{G}\}$
2. **Logit Transformation**: Convert probabilities to logits to handle extreme values and provide unbounded feature space:
$$\mathbf{L}_a = \log(\text{clip}(\mathbf{P}_a, \epsilon, 1.0)) \quad (1)$$
where $\epsilon = 10^{-6}$ prevents numerical instability.
3. **Concatenation**: For each base model, concatenate aspect logits:
$$\mathbf{X}_{\text{base}} = [\mathbf{L}_{\mathrm{E}} \,||\, \mathbf{L}_{\mathrm{S}} \,||\, \mathbf{L}_{\mathrm{G}}] \in \mathbb{R}^{n \times 12} \quad (2)$$

This transformation preserves relative probability magnitudes while providing a more stable feature space for meta-learning, avoiding the compression of probabilities near 0 or 1 that can occur in linear scaling.

4.5. Meta-Level Ensemble Architecture

We propose two hierarchical ensemble strategies that differ in their aggregation topology.

Tower A employs a single-stage meta-learner that processes concatenated meta-features from all selected base families:

$$\mathbf{X}_{\text{meta}} = [\mathbf{X}_{\text{fam}_1} \,||\, \mathbf{X}_{\text{fam}_2} \,||\, \cdots \,||\, \mathbf{X}_{\text{fam}_k}] \in \mathbb{R}^{n \times 12k} \tag{3}$$

where k is the number of base model families. This architecture allows the meta-learner to discover arbitrary cross-family interaction patterns.

Tower B implements a two-level hierarchy to exploit family-specific characteristics:

1. **Level 1 (Family-Specific Meta-Models)**: Each base family i trains an independent meta-MLP:

$$\mathbf{Z}_i = \text{MLP}_{\text{fam}_i}(\mathbf{X}_{\text{fam}_i}) \in \mathbb{R}^{n \times 12} \tag{4}$$

2. **Level 2 (Cross-Family Aggregation)**: A second meta-MLP combines family-level outputs:

$$\hat{\mathbf{Y}} = \text{MLP}_{\text{final}}([\mathbf{Z}_1 \,||\, \mathbf{Z}_2 \,||\, \cdots \,||\, \mathbf{Z}_k]) \tag{5}$$

This architecture allows each family to learn specialized combination strategies (e.g., TabPFN families may benefit from uncertainty calibration while transformer families may require confidence rescaling) before global aggregation.

Meta-MLP Architecture All meta-models share a unified neural architecture with multi-task learning formulation:

$$\mathbf{h}^{(1)} = \text{ReLU}(\text{BN}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)})) \tag{6}$$

$$\mathbf{h}^{(1)}_{\text{drop}} = \text{Dropout}(\mathbf{h}^{(1)}, p = 0.4) \tag{7}$$

$$\mathbf{z} = \text{ReLU}(\mathbf{W}^{(2)}\mathbf{h}^{(1)}_{\text{drop}} + \mathbf{b}^{(2)}) \tag{8}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{64 \times d_{\text{in}}}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{64 \times 64}$, and BN denotes batch normalization. The shared trunk $\mathbf{z}$ feeds into three aspect-specific prediction heads:

$$\mathbf{o}_a = \mathbf{W}_a^{(3)}\text{ReLU}(\mathbf{W}_a^{(2)}\mathbf{z} + \mathbf{b}_a^{(2)}) + \mathbf{b}_a^{(3)} \tag{9}$$

for $a \in \{\text{E}, \text{S}, \text{G}\}$, where $\mathbf{W}_a^{(2)} \in \mathbb{R}^{32 \times 64}$ and $\mathbf{W}_a^{(3)} \in \mathbb{R}^{4 \times 32}$.

The multi-task formulation with shared representations encourages learning of correlations between ESG aspects (e.g., environmental practices often correlate with governance structures) while maintaining aspect-specific prediction capacity.

Loss Function: Joint cross-entropy across all aspects:

$$\mathcal{L} = \text{CE}(\mathbf{o}_{\text{E}}, \mathbf{y}_{\text{E}}) + \text{CE}(\mathbf{o}_{\text{S}}, \mathbf{y}_{\text{S}}) + \text{CE}(\mathbf{o}_{\text{G}}, \mathbf{y}_{\text{G}}) \tag{10}$$

Optimization: AdamW with learning rate 10^{-3}, weight decay 0.01, batch size 64.

4.5.1. Training Protocol: Stratified 80/20 Split

We employ a stratified holdout protocol to balance computational efficiency with robust evaluation. The training procedure consists of three stages:

Stage 1: Data Partitioning The training corpus $\mathcal{D}_{\text{train}}$ is partitioned into training ($\mathcal{D}_{80}$) and validation ($\mathcal{D}_{20}$) subsets using stratified sampling. To preserve the joint distribution of ESG labels, we implement multilabel stratification on the (E, S, G) triplets using iterative stratification (Sechidis et al., 2011). This ensures that the validation set maintains representative samples from all 64 possible ESG label combinations ($4 \times 4 \times 4$), preventing evaluation bias from rare triplet configurations.

Stage 2: Base Model Training

Each base model family is trained exclusively on $\mathcal{D}_{80}$ and generates predictions on both $\mathcal{D}_{20}$ (validation) and $\mathcal{D}_{\text{test}}$ (held-out test set):

1. **Embedding Models + TabPFN**: Extract embeddings from $\mathcal{D}_{80}$, optionally apply SVD dimensionality reduction, fit TabPFN classifier, predict on $\mathcal{D}_{20}$ and $\mathcal{D}_{\text{test}}$

2. **Transformer Models**: Fine-tune on $\mathcal{D}_{80}$ with early stopping based on $\mathcal{D}_{20}$ performance, generate final predictions on $\mathcal{D}_{20}$ and $\mathcal{D}_{\text{test}}$ using best checkpoint

This produces two sets of meta-features per base model:

- $\mathbf{X}_{\text{meta}}^{(20)} \in \mathbb{R}^{|\mathcal{D}_{20}| \times 12}$: Meta-features for validation samples
- $\mathbf{X}_{\text{meta}}^{(\text{test})} \in \mathbb{R}^{|\mathcal{D}_{\text{test}}| \times 12}$: Meta-features for test samples

Stage 3: Meta-Model Training. Meta-models are trained on $\mathbf{X}\text{meta}^{(20)}$ with early stopping: split $\mathcal{D}20$ into $80\%/20\%$ (stratified) meta-train/validation, train up to 200 epochs while tracking validation loss, select t^* with minimum validation loss (patience=15), retrain on all of $\mathcal{D}_{20}$ for t^* epochs, then generate final predictions on $\mathcal{D}_{\text{test}}$. This nested validation promotes generalization and reduces overfitting to base-model biases.

To ensure robustness against random initialization effects, we repeat the entire pipeline across three independent random seeds $\mathcal{S} = \{0, 100, 200\}$.

4.6. Evaluation Metrics

Model performance is assessed using four complementary metrics, computed independently for each ESG aspect:

- **Accuracy**: $\text{Acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\hat{y}_i = y_i]$

- **Macro-averaged F1**: Harmonic mean of precision and recall across classes without class-weighting:

$$\text{F1}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\cdot \text{Prec}_c \cdot \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c} \tag{11}$$

- **Balanced Accuracy**: Arithmetic mean of per-class recall, accounting for class imbalance:

$$\text{BAcc} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{12}$$

- **Area Under Precision-Recall Curve (AUPRC)**: Average precision across one-vs-rest binary decompositions, providing a single-number summary of the precision-recall trade-off

System-level performance is reported as the mean across aspects and seeds, with standard deviation indicating inter-seed variability.

5. Results and Discussion

The results presented in Tables 3–5 provide clear evidence that transformer-based architectures, supported by ensemble and multi-task learning strategies, are well suited for ESG sentiment classification in Slovene news. The consistent performance gains achieved by the multi-task fusion models across all three ESG dimensions indicate that the aspects of Environmental, Social, and Governance sentiment share underlying linguistic cues that can be effectively captured through shared representations. This interdependence highlights that public discourse around ESG topics is often contextually entangled—positive environmental narratives tend to correlate with favorable governance and social framing, and vice versa.

The Environmental aspect (Table 3) shows the strongest overall results, with macro-F1 values surpassing 0.6 for the top-performing models. The superior accuracy of the ensemble "Final Tower" architectures suggests that aggregating diverse feature spaces—sentence embeddings, fine-tuned transformer outputs, and meta-learned representations—yields a more comprehensive understanding of ESG-related sentiment. Notably, the SloBERTa model achieves robust scores comparable to or exceeding multilingual alternatives, confirming that monolingual pretraining remains advantageous for highly inflected languages such as Slovene. By contrast, multilingual models like XLM-RoBERTa exhibit more stable but less specialized behavior, implying that language-agnostic pretraining can miss subtle morphological or idiomatic sentiment signals present in the Slovene media corpus. Performance for the Social aspect (Table 4) is comparatively lower, with macro-F1 values clustering between 0.30 and 0.45 across models. This can be attributed to the abstract and context-sensitive nature of social issues—topics like labor relations or equality often rely on nuanced framing rather than explicit sentiment markers. Interestingly, the multilingual models performed competitively in this category, suggesting that cross-lingual exposure may help recognize generalized social discourse patterns. The relative underperformance of ensemble systems in this dimension further supports the idea that social sentiment requires more contextual or pragmatic interpretation than currently encoded by the models. The Governance aspect (Table 5) remains the most challenging dimension, reflected in lower average macro-F1 values and wider variance between seeds. This weakness likely stems from ambiguity in annotator interpretations and the abstract, institutional tone typical of governance reporting. Governance language often lacks clear evaluative expressions, making sentiment polarity difficult to infer even for human annotators. The observed correspondence between lower inter-annotator agreement and reduced model performance supports this interpretation and underlines the difficulty of operationalizing governance sentiment in textual data.

Overall, the results validate the study's design choices while also exposing limitations inherent to ESG text analysis in news. The small dataset size (550 annotated articles) constrains generalization, particularly for multi-class classification across three interrelated sentiment axes. Moreover, the reliance on LLM-assisted filtering introduces potential sampling bias, as model-based preselection may favor easily classifiable or lexically explicit texts. The ESG sentiment categories themselves may overlap semantically, challenging both human and machine annotation consistency. Future studies could mitigate these issues through larger, more balanced datasets and clearer annotation guidelines emphasizing cross-aspect distinctions.

6. Case Study: Qualitative Temporal ESG Evaluation

After evaluating the proposed models, we select the gpt-oss-20b model to analyze the sentiment distribution over time for four companies of interest by analysing a large news media monitoring dataset for the period 2010-2025. The annual average sentiment score is computed by subtracting the count of negative sentiment articles from the count of positive sentiment articles.

These companies were selected as representative cases of different approaches to sustainability,

Table 3: Test-set results for Aspect E (mean over seeds). Primary metric is F1-macro; we also report AUPRC, Balanced Accuracy (BAcc), and Accuracy. Best results per column are in bold.

Model	Accuracy	F1-macro	BAcc	AUPRC
Baseline				
Majority	0.7000	0.2059	0.2500	0.2500
Ensemble				
FinalTowerA	0.7394	0.4469	0.5010	0.5638
FinalTowerB	0.7182	0.4291	0.4476	0.4815
Fine-tuned Transformers				
hf_sloberta	0.7242	0.4284	0.4648	0.5166
hf_xlm-roberta	0.7212	0.4251	0.4572	0.5260
Sentence-Transformer				
tabpfn_bge-m3	0.7455	0.3841	0.3986	0.5198
tabpfn_gemma-embed	0.7182	0.3402	0.3412	0.4404
tabpfn_paraphrase	0.7091	0.3691	0.3972	0.4548
SVD				
tabpfn_bge-m3_svd	0.7727	0.4717	0.4648	0.5847
tabpfn_gemma-embed_svd	0.7212	0.3972	0.3995	0.4894
tabpfn_paraphrase_svd	0.7424	0.3970	0.4103	0.5107
LLMs				
GaMS-27B 10	0.8000	0.5108	0.5102	0.4441
GaMS-9B	0.7545	0.4255	0.3926	0.3484
Gemma3-12B	0.7818	0.5375	0.5449	0.4546
Gemma3-27B	0.7818	**0.6106**	**0.6777**	0.4895
gpt-oss 20B	**0.8182**	0.5907	0.5828	**0.5847**

Table 4: Test-set results for Aspect S (mean over seeds). Primary metric is F1-macro; we also report AUPRC, Balanced Accuracy (BAcc), and Accuracy. Best results per column are in bold.

Model	Accuracy	F1-macro	BAcc	AUPRC
Baseline				
Majority	0.3364	0.1259	0.2500	0.2500
Ensemble				
FinalTowerA	0.4606	0.3033	0.3589	0.3515
FinalTowerB	0.4364	0.3252	0.3602	0.3557
Fine-tuned Transformers				
hf_sloberta	0.4909	0.4238	0.4265	0.4475
hf_xlm-roberta	0.4939	0.4404	0.4423	**0.4533**
Sentence-Transformer				
tabpfn_bge-m3	0.4455	0.2726	0.3360	0.4051
tabpfn_gemma-embed	0.4515	0.3177	0.3529	0.4083
tabpfn_paraphrase	0.4424	0.2674	0.3336	0.4072
SVD				
tabpfn_bge-m3_svd	0.4515	0.2804	0.3418	0.3821
tabpfn_gemma-embed_svd	0.4727	0.2940	0.3600	0.3962
tabpfn_paraphrase_svd	0.4545	0.2804	0.3443	0.3987
LLMs				
GaMS-27B	0.5000	0.4140	0.4358	0.3395
GaMS-9B	0.5182	0.4193	0.4432	0.3529
Gemma3-12B	0.4182	0.3717	0.4179	0.3369
Gemma3-27B	0.4455	0.4022	**0.4622**	0.3430
gpt-oss 20B	**0.5273**	**0.4512**	0.4547	0.3603

governance, and community engagement within Slovenian industry. *Talum* exemplifies a successful transition from a high–environmental-impact aluminum producer to a recycling-based model, effectively balancing environmental responsibility with its role as a major regional employer. Similarly, *SDH*, as a state holding company managing publicly owned enterprises, has introduced high governance standards and driven the adoption of ESG reporting across state-managed firms, consistent with research suggesting that government ownership often fosters sustainability commitments (Qian and Yang, 2023).

Table 5: Test-set results for Aspect G (mean over seeds). Primary metric is F1-macro; we also report AUPRC, Balanced Accuracy (BAcc), and Accuracy. Best results per column are in bold.

Model	Accuracy	F1-macro	BAcc	AUPRC
Baseline				
Majority	0.4818	0.1626	0.2500	0.2500
Ensemble				
FinalTowerA	0.5152	0.3742	0.3997	0.4351
FinalTowerB	0.4939	0.3935	0.4158	0.4342
Fine-tuned Transformers				
hf_sloberta	0.6091	**0.5420**	**0.5486**	**0.5528**
hf_xlm-roberta	0.5333	0.3590	0.3843	0.4849
Sentence-Transformer				
tabpfn_bge-m3	0.5879	0.4457	0.4571	0.5164
tabpfn_gemma-embed	0.5636	0.3867	0.3974	0.4751
tabpfn_paraphrase	0.5818	0.4239	0.4337	0.4739
SVD				
tabpfn_bge-m3_svd	**0.6879**	0.5210	0.5441	0.5386
tabpfn_gemma-embed_svd	0.6030	0.3909	0.4248	0.4867
tabpfn_paraphrase_svd	0.5788	0.4119	0.4349	0.4560
LLMs				
GaMS-27B	0.5000	0.3899	0.4405	0.3620
GaMS-9B	0.6000	0.4452	0.4591	0.3730
Gemma3-12B	0.4727	0.4870	0.5120	0.4294
Gemma3-27B	0.4091	0.4146	0.4529	0.3816
gpt-oss 20B	0.5364	0.4792	0.4821	0.3839

Table 6: ESG Sentiment Analysis Summary by Company and Category

Company	Category	Total	Relevant	Positive	Negative	Neutral	Irrelevant
Talum	E	4072	1234	460	448	326	2838
	S		2446	952	900	594	1626
	G		2404	512	1080	812	1668
Sdh	E	30338	2504	668	838	998	27834
	S		15758	1892	7908	5958	14580
	G		27640	2082	14610	10948	2698
Cinkarna	E	11062	1516	364	794	358	9546
	S		2754	666	1178	910	8308
	G		3502	418	1696	1388	7560
Salonit	E	8026	1408	356	862	190	6618
	S		1810	464	970	376	6216
	G		1970	234	1182	554	6056

Cinkarna Celje reflects a long-term transformation from a historically polluting zinc producer to a company committed to environmental remediation and local well-being, actively monitoring soil conditions and the health of nearby residents. In contrast, *Anhovo / Alpacem* illustrates the social tensions that can arise when industrial development and community interests diverge. Once associated with asbestos production and its severe societal impacts, the company's more recent plans to expand into waste incineration have sparked public resistance due to uncertainties about environmental and health consequences. Together, the cases capture a spectrum of corporate responses to sustainability pressures—from proactive adaptation and transparency to ongoing conflict and mistrust.

In this case study, an economic expert compared temporal ESG-sentiment analysis with key business events as described in relevant CEO letters for Cinkarna Celje and Talum.

At Cinkarna Celje, ESG sentiment patterns indicate strong sensitivity to regulatory and

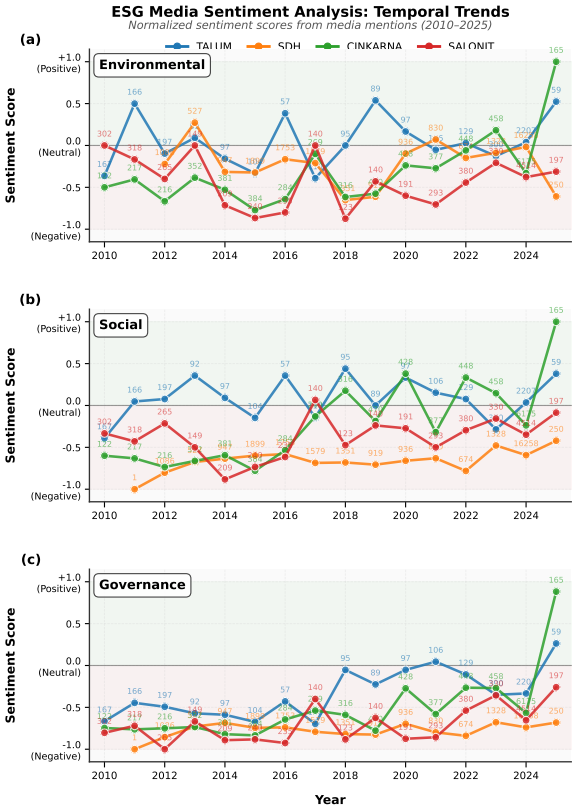


Figure 3: Normalized sentiment scores from media mentions (2010 2025)

governance-related events. The announcement of environmental remediation in 2017 led to a significant improvement in E and S sentiment, while the subsequent slow remediation process, coupled with a lawsuit by the European Commission over delays in closing the landfills and the question at the EU level regarding whether the raw material for core product, titanium dioxide, is carcinogenic, affected sentiment. G sentiment increased during board and management changes in 2020 and again in 2025 with the reappointment of the same CEO, but declined in 2024 when a board member resigned, suggesting that leadership stability and continuity are positively valued. Governance sentiment also correlates positively with dividend payments, reflecting an association between shareholder returns and perceptions of governance quality. The exceptionally high sentiment in 2025 coincides with the European Court of Justice ruling that titanium dioxide is not carcinogenic – a decision with limited impact on actual environmental and health outcomes, but significant reputational impact, illustrating how institutional signals can reshape ESG perception independently of environmental performance. At Talum in 2015, financial results turned positive after a prolonged period of losses: strategic goals were achieved, the main shareholder increased its equity investment, and employees were highly engaged in innovative processes. However, sentiment in all three pillars (E, S, and G) declined, suggesting persistent scepticism despite improved financial performance. Sentiment rebounded in 2016, supported by strategic restructuring, innovation, and workforce expansion, but fell again in 2017 as environmental sentiment weakened despite the company's continued commitment to efficient and sustainable production, an increased workforce, and doubled profit. Between 2018 and 2019, E sentiment strengthened as Talum invested in restructuring its production towards carbon-neutral products with high added value, while S sentiment declined due to perceived risks to employment associated with the reduction of primary aluminium production. In 2025, all three sentiment dimensions were strongly positive, reflecting the completion of Talum's green transformation, technological modernisation, and diversification into new industries such as commerce, pharmaceuticals, and defence. It is notable that investments in defence seem no longer to be considered ESG "problematic" in 2025, probably due to the geopolitical situation. Both firms exhibited markedly positive S sentiment in 2020, coinciding with the COVID-19 pandemic. This increase is likely related to the companies' ability to maintain stable operations and retain employees despite disrupted market conditions, reinforcing employment security as a key driver of social sentiment. Across both companies, announcements and disbursements of employee bonuses consistently coincided with positive shifts in S sentiment, suggesting that distributive and welfare-related actions have a measurable influence on social evaluations. Workforce contraction had limited effect on S sentiment in Cinkarna, whereas in Talum, S sentiment is highly sensitive to any possible impact of any of the conditions on employment. Across the two companies, ESG sentiment is only weakly related to financial indicators (profit, liquidity, efficiency). Instead, communicative and institutional factors - regulatory decisions, board changes, and employee-related gestures - exert a stronger and more immediate effect. These findings indicate that text-based ESG sentiment primarily reflects the social construction of corporate responsibility rather than direct economic or environmental outcomes.

7. Conclusions and Further Work

This work presents the first publicly available Slovene ESG dataset and uses it as a resource for training LLM-based, transformer-based classification models and hierarchical ensembling. Beyond technical performance, these findings have broader implications for sustainability analytics: automated monitoring of ESG sentiment could provide dynamic, fine-grained insights into corporate reputation shifts across time and media outlets.

Our results show that LLMs lead Environmental (Gemma3-27B, F1-macro 0.61) and Social (gpt-oss 20B, 0.45) tasks, while fine-tuned SloBERTa tops Governance (0.54). Future research should pursue several directions. Expanding the dataset temporally and thematically would enhance robustness. Incorporating temporal and causal modeling could capture how specific events—policy changes, environmental incidents, or governance scandals—affect sentiment trajectories. The most interesting line of research is to observe the ESG assigned sentiment in relation to ESG financial information.

8. Code Availability

The source code is publicly available at https://github.com/bkolosk1/slo-news-esg.

9. Data Availability

The dataset is publicly available at http://hdl.handle.net/11356/2102.

Acknowledgments

This work was supported by the Slovenian Research and Innovation Agency (ARIS) through the projects EMMA (Embeddings-based Techniques for Media Monitoring Applications; L2-50070), Large Language Models for Digital Humanities (LLM4DH; GC-0002), and the research core funding programme Knowledge Technologies (P2-0103). BK is supported by the Young Researcher Grant PR-12394.

10. Bibliographical References

S. Angioni et al. 2024. Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis. *IEEE Access*.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 66–71. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Ebrahim Bazrafshan. 2023. The role of ESG ranking in retail and institutional investors' attention and trading behavior. *Finance Research Letters*, 58:104462.

Simin Chen, Yu Song, and Peng Gao. 2023. Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance. *Journal of environmental management*, 345:118829.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186. Association for Computational Linguistics.

Gregor Dorfleitner and Jun Zhang. 2024. Esg news sentiment and stock price reactions: A comprehensive investigation via bert. *Schmalenbach Journal of Business Research*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Gemma Team. 2025. Gemma 3 technical report.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. TabPFN: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Nikola Ivačič, Thi Hong Hanh Tran, Boshko Koloski, Senja Pollak, and Matthew Purver. 2023. Analysis of transfer learning for named entity recognition in South-Slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language*

Processing 2023 (SlavicNLP 2023), pages 106–112, Dubrovnik, Croatia. Association for Computational Linguistics.

Boshko Koloski, Syrielle Montariol, Matthew Purver, and Senja Pollak. 2022. Knowledge informed sustainability detection from short financial texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taja Kuzman and Nikola Ljubešić. 2025. LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access*.

H. Lee, J. H. Kim, et al. 2024. Deep-learning-based stock market prediction incorporating esg sentiment and technical indicators. *Scientific Reports*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David C.L. Ngo. 2015. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

OpenAI Team. 2025. gpt-oss-120b & gpt-oss-20b model card.

Ting Qian and Caoyuan Yang. 2023. State-owned equity participation and corporations' ESG performance in China: The mediating role of top management incentives. *Sustainability*, 15(15).

Tobias Schimanski et al. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of CIKM*.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual BERT-based language model. In *Text, Speech and Dialogue (TSD 2021)*, Cham. Springer.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik Šikonja. 2024. *Generative model for less-resourced language with 1 billion parameters*, page 485–511.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Not All News Is Equal: Topic- and Event-Conditional Sentiment from Finetuned LLMs for Aluminum Price Forecasting

Alvaro Paredes Amorin[1,2], **Andre Python**[2,3,4], **Christoph Weisser**[5]
[1] International Business School, Zhejiang University
[2] Center for Data Science, Zhejiang University
[3] Centre for Human Genetics, Nuffield Department of Medicine, Oxford University
[4] School of Medicine, Zhejiang University
[5] Bielefeld School of Business, Hochschule Bielefeld (HSBI) - University of Applied Sciences and Arts
alvaroparedesamorin@gmail.com python.andre@gmail.com christoph.weisser@hsbi.de

Abstract

By capturing the prevailing sentiment and market mood, textual data has become increasingly vital for forecasting commodity prices, particularly in metal markets. However, the effectiveness of lightweight, finetuned large language models (LLMs) in extracting predictive signals for aluminum prices—and the specific market conditions under which these signals are most informative—remains under-explored. This study generates monthly sentiment scores from English and Chinese news headlines (Reuters, Dow Jones Newswires, and China News Service) and integrates them with traditional tabular data, including base metal indices, exchange rates, inflation rates, and energy prices. We evaluate the predictive performance and economic utility of these models through long-short simulations on the Shanghai Metal Exchange from 2007 to 2024. Our results demonstrate that during periods of high volatility, Long Short-Term Memory (LSTM) models incorporating sentiment data from a finetuned Qwen3 model (Sharpe ratio 1.04) significantly outperform baseline models using tabular data alone (Sharpe ratio 0.23). Subsequent analysis elucidates the nuanced roles of news sources, topics, and event types in aluminum price forecasting

Keywords: aluminum price, natural language processing, large language model, sentiment analysis, finance

1. Introduction

Aluminum is a key non-ferrous metal with widespread applications in automotive, aerospace, construction, and electronics as a result of its lightweight, high corrosion resistance, and excellent conductivity. Aluminum production represents approximately 3.5% of the electricity consumed worldwide and contributes to approximately 1% of global carbon emissions, making it a highly energy-intensive and strategically important commodity (Cullen and Allwood, 2013; Yi et al., 2024; International Aluminium Institute, 2021). Its price dynamics is influenced by a combination of supply constraints, energy costs, geopolitical developments, and demand from emerging industries such as electric vehicles and renewable energy infrastructure (Luglio, 2023; Bastin, 2024). These factors contribute to an increase in price volatility, which presents challenges for both market participants and industrial decision-makers.

Aluminum price forecasting typically uses statistical and machine learning methods that relies on tabular (numerical) data on historical prices within a time series modeling framework (Sverdrup et al., 2015; Esangbedo et al., 2024; Oikonomou and Damigos, 2024). Although methods that exclusively use tabular data can capture some patterns in price variations, they cannot capture information from textual data sources, such as news headlines and analyst reports, which can provide complementary information that reflect, e.g., investor sentiment and expectations.

We investigate whether the sentiment derived from finetuned large language models (LLMs) can improve aluminum price prediction and inform trading strategies. Specifically, we construct sentiment variables from English and Chinese news headlines, classify them with finetuned LLMs, and integrate these signals with numerical time-series data to forecast monthly aluminum prices. Our study also assesses how news sources, topics, and event types can lead to variation in the relevance of signals. By combining time-series models with finetuned LLM sentiment, we identify the conditions under which textual information provides the greatest economic value, offering both methodological insights and practical guidance for commodity market forecasting.

2. Related Work

2.1. Statistical and machine learning approaches to forecast metal prices

Non-ferrous prices have been commonly forecasted using statistical models applied to time series data, such as Runge Kutta methods (Sverdrup et al., 2015) and autoregressive integrated moving average (ARIMA) models (Dooley and Lenihan, 2005; Kriechbaumer et al., 2024). More recently, Mysen and Thornton (2021) showed that tree-

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 59–77

based algorithms, such as extreme gradient boosting (XGBoost), can outperform statistical methods in predicting aluminum prices. Oikonomou and Damigos (2024) showed that auto-regressive light gradient-boosting machine models can further improve the predictive performance on aluminum returns over six months. Larger models based on recurrent neural network architectures, such as long-short term memory (LSTM) models (Hochreiter and Schmidhuber, 1997)—a type of neural network model particularly effective at capturing long-term dependencies within time series data (Huang, 2024)—have shown promising results in predicting aluminum prices (Esangbedo et al., 2024). Both models integrating several machine learning techniques, also-called hybrid models (Li et al., 2023), and models aggregating the outputs of multiple models, which refer to ensemble models (Esangbedo et al., 2024), showed promising results in the forecast of aluminum and other nonferrous metal prices.

2.2. Role of textual data in stock price forecast

Understanding market sentiment can provide a valuable context for interpreting analyst forecasts Chen et al. (2020). For example, Kumar and Ravi (2021); Thormann et al. (2021); Kant et al. (2024) find that the integration of sentiment scores with traditional financial indicators can improve stock price forecasts. When combined with data on copper and aluminum commodities prices, Sinatrya et al. (2022) found that sentiment analysis can help predict the value of metal industry companies on the stock market. While Chen et al. (2021) showed that social media sentiment improves the accuracy of traditional econometric models for copper price forecasting, Gupta et al. (2020) found that sentiments derived from Twitter and news articles positively correlates with short-term prices of gold and silver. Within a hybrid model framework Yuan et al. (2020) added a module on the opinion score from a Chinese news website that improved the predictive accuracy of short-term gold prices.

2.3. A growing influence of large language models

Large language models (LLMs) play an increasing role in the prediction of financial assets. Without the need for explicit rules, LLMs can capture sentiments from various languages by learning contextual representations via pre-training on massive corpora.

Several studies demonstrate that finetuned LLM-based sentiment signals outperform traditional financial NLP benchmarks when used in forecasting tasks. Kumar and Singh (2024) show that an LLaMA3 model finetuned on FinancialPhraseBank leads to a superior sentiment classification performance compared to the base LLaMA3 model. Zhang et al. (2023a) further report that instruction-tuned LLaMA 7B models provide more informative signals for downstream financial prediction tasks than non-instructed variants. Similarly, Zhang et al. (2023b) find that combining fine-tuning with retrieval-augmented generation improves the economic relevance of model outputs. Applying this to a portfolio comprised of 417 stocks from the S&P 500, Konstantinidis et al. (2024) obtained the highest returns with their finetuned LLaMA2 model, outperforming other lexicon-based methods and FinBERT.

The effectiveness of adaptation techniques is also observed across learning regimes. Fatemi et al. (2024) show that fine-tuning Flan-T5 models significantly improves forecasting-related performance compared to zero- and few-shot setups, while adapted open-source LLMs such as LLaMA3, Mistral and Phi consistently outperform finance baseline models. Comparable gains from fine-tuning are reported by Wang et al. (2023) in multiple open-source LLMs, with LLaMA2 and MPT producing the strongest downstream results.

Finetuned LLMs can efficiently capture market-relevant information not only from English but also from, e.g., Chinese textual data (Lan et al., 2023). More recently, Paredes Amorin et al. (2025) show that DeepSeek, Qwen, and LLaMA models fine-tuned on English and Chinese languages can consistently outperform the benchmark financial model FinBERT, which is a pre-trained BERT-based language model specifically designed to analyze sentiment and extract information from financial text.

3. Methods

To assess the role of sentiment from news data in predicting closing aluminum prices, we compare several time series forecasting models that use tabular ("Data 1") and/or sentiment ("Data 2") data. The general workflow is summarized in Figure 1.

3.1. Data

This research considers tabular (numerical) data gathered from the Wind terminal of the Shanghai Stock Exchange price index. It includes daily closing prices for aluminum using the Aluminum ingots commodity and factors identified as key drivers of aluminum prices (Esangbedo et al., 2024). This includes the exchange rates of the Chinese and US currencies, the US and China inflation rates, and the closing prices of copper, zinc, and iron, as well as the prices of crude oil (OIL), Brent crude oil (LCOU5) and natural gas (NGQ5). These nu-

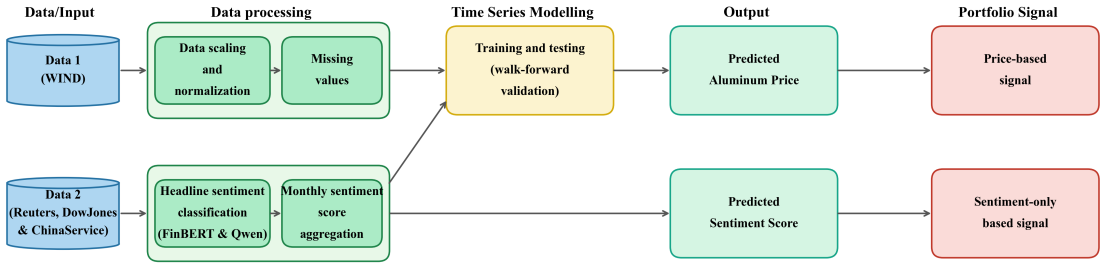


Figure 1: **Workflow.** *Data 1*: financial data from WIND terminal includes tabular data extracted from March 2007 to April 2024 (4,152 rows). *Data 2*: textual data that includes headlines from two news sources in English (Reuters (N=4,963), Dow Jones Newswires (N=11,581), and a news source in Chinese (China News Service (N=8,970)) collected from March 2007 to April 2024. The data processing and sentiment analysis (green) includes data scaling, normalization, and the treatment of missing values for the tabular data, and the use of language models to generate new sentiment variables from *Data 2*. The sentiment is classified in "positive", "negative" or "neutral". Monthly sentiment scores are combined with other numerical data (yellow box) to train and test time series models in order to predict monthly aluminum prices.

merical daily data are then aggregated into monthly values during the investigated period, from March 2007 to April 2024, in line with the temporal granularity of the textual news datasets. We collected textual data related to aluminum from 2007 to 2024 from Factiva, a business information and research platform that contains a large news database. The dataset was filtered to remove noise and eliminate entries lacking relevance to aluminum price dynamics. The filtering method is described in Appendix A. We focus on headlines (final number of headlines in parentheses)—it is common among NLP and sentiment studies to use headlines instead of full news articles (Ewald and Li, 2024; Breitung et al., 2023)—in English, from Reuters (N=4,963) and Dow Jones Newswires (N=11,581), and one dataset in Chinese (mandarin) from the China News Service (N=8,970) datasets. The data used in this study from Factiva cannot be redistributed due to licensing restrictions.

3.2. Models

To classify aluminum news datasets, we investigate FinBERT and a lightweight LLM (Qwen3) finetuned by Paredes Amorin et al. (2025)[1] with five different financial sentiment datasets: FinancialPhraseBank, Financial Question Answering, Gold News Sentiment, Twitter Sentiment and Chinese Finance Sentiment. These datasets cover a diverse range of financial text sources, including expert-annotated news sentences, financial document question-and-answer pairs, commodity related news, social media content, and Chinese language financial news. Here, the LLMs use the aluminum-related news to predict the sentiment of each news' headline. The models classify sentiment in three categories: "positive" = +1, "negative" = -1 & "neutral" = 0. When using the finetuned Qwen model, we use the same prompt as in the finetuning process with financial sentiment datasets, and it returns one of the three labels. In the case of FinBERT, it is a BERT classifier model trained on financial data and it outputs the most likely label among the three categories by default. We computed a weighted sum of each score associated with a news item in a given month to account for the presence of multiple news items per month. For example, a score of 0.4 is given for 6 positive, 2 neutral and 2 negative news $(6 \cdot 1 + 2 \cdot 0 + (-1) \cdot 2)/10 = 0.4)$.

[1]The source code and experimental pipeline are publicly available at Github

3.3. Defining trading signal by trading strategy

To compare sentiment-only based and price-based strategies to forecast aluminum prices, we define trading signals adapted to each strategy from which we can compare the use of news sentiment scores alone compared to the use of prices predicted by time series forecasting models using numerical (tabular) data, including sentiment scores. We define Sent_t as the average sentiment score of all news headlines published during period t. We consider a sentiment-only based trading strategy with associated trading sentiment signal Ss_t as follows:

$$Ss_t = \begin{cases} +1 & \text{if } \text{Sent}_t > 0 \quad \text{(buy/long)} \\ -1 & \text{if } \text{Sent}_t < 0 \quad \text{(sell/short)} \\ 0 & \text{if } \text{Sent}_t = 0 \quad \text{(neutral)} \end{cases} \tag{1}$$

Next, we implement a price-based trading strategy using the optimal time-series forecasting model. This model was selected from various configurations via the methodology detailed in Appendix B, with performance results provided in Appendix C. Let P_t^{true} represent the actual aluminum price at time t, and P_{t+1}^{pred} represent the predicted price for the following period. The resulting trading signal, Sn_t, is defined as:

$$Sn_t = \begin{cases} +1 & \text{if } P_{t+1}^{\text{pred}} > P_t^{\text{true}} \quad \text{(buy/long)} \\ -1 & \text{if } P_{t+1}^{\text{pred}} < P_t^{\text{true}} \quad \text{(sell/short)} \\ 0 & \text{otherwise} \quad \text{(neutral)} \end{cases} \tag{2}$$

To compute portfolio performance, the generated signals Ss_t and $Sn_t \in \{-1, 0, +1\}$ are multiplied by the aluminum return realized for each period.

3.4. Evaluation Metrics

We evaluate the economic performance of the proposed trading strategies using cumulative return and the Sharpe ratio.

Let R_t denote the simple monthly return at time t. The cumulative return over the evaluation period $[0, T]$ is computed as:

$$R_{0,T}^{\text{cum}} = \prod_{t=1}^{T}(1 + R_t) - 1 \tag{3}$$

This metric measures the total compounded growth of the strategy over the full sample period and directly reflects long-term investment performance.

To assess risk-adjusted performance, we compute the Sharpe ratio defined as:

$$\text{SR} = \frac{\bar{R} - R_f}{s} \tag{4}$$

where $\bar{R}$ is the average monthly return, R_f is the monthly risk-free rate, and s is the standard deviation of monthly returns. The Sharpe ratio captures the excess return per unit of risk, allowing comparison across competing forecasting models and trading strategies.

4. Results and discussion

4.1. Role of trading strategy and market volatility

We assess the model performance by trading strategy (tabular-only, tabular+sentiment, sentiment only (Qwen sentiment), sentiment only (FinBERT sentiment)) and three volatility scenarios. We partition the sample into three volatility scenarios based on a 6-month rolling standard deviation of aluminum returns, annualized by multiplying by $\sqrt{12}$. We define volatility scenarios based on fixed percentages of the observed volatility range to ensure economically meaningful thresholds. Specifically, the 20% less volatile is considered as a low-volatility range, while the 20-50% is considered medium-volatility and higher than the 50% is high-volatility. In our sample, annualized volatility ranges from 1.13% to 37.47%, yielding threshold values of 8.40% and 19.30% for the low-medium and medium-high boundaries, respectively. This classification produces 66 low-volatility months, 106 medium-volatility months, and 28 high-volatility months. Appendix D shows the resulting volatility regimes distributed across time (Panel A) as well as portfolio value of the different strategies (Panel B).

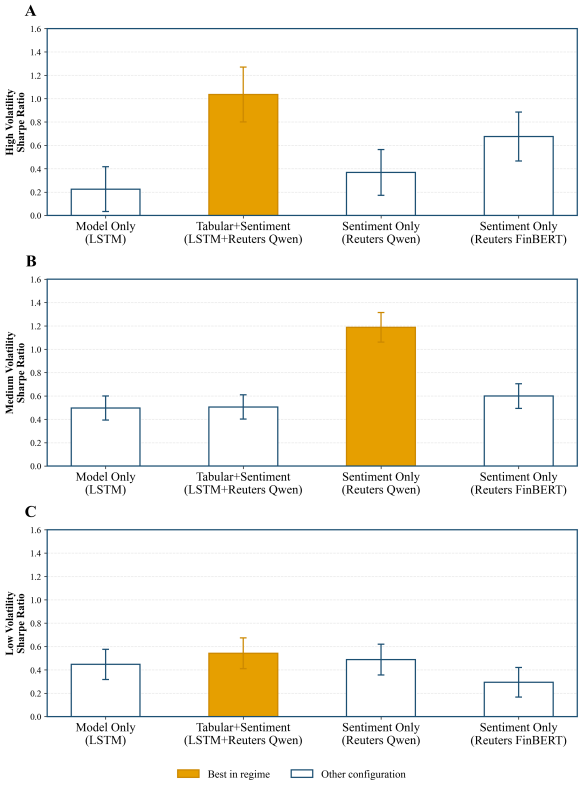


Figure 2: **Evaluation of portfolio's performance by strategy and volatility scenarios.** For each strategy (tabular-only, tabular+sentiment, sentiment only (qwen), sentiment only (reuters)) the portfolio's performance is represented by the Sharpe ratio (y-axis) across three volatility scenarios (panels), with: **A** high volatility scenario (n=28 months), **B** medium volatility (n=106 months), and **C** low volatility (n=66 months). Error bars represent ±1 standard error. The highlighted bars (orange) indicate the best performing strategy within each scenario.

Figure 2 shows the estimated Sharpe ratio for

each strategy across three volatility regimes (panels A–C). During high-volatility periods, the integrated tabular and sentiment strategy achieves the highest Sharpe ratio (1.04), a 359% improvement over the tabular-only baseline (0.23). This suggests that in turbulent markets—where historical price correlations often deviate—news sentiment captures critical directional signals, such as panic or recovery, which traditional time-series models fail to incorporate. In medium-volatility periods, the sentiment-only strategy dominates all others (Sharpe ratio = 1.19), substantially outperforming even the combined approach (0.51). This suggests that, under normal market conditions, the sentiment signal is sufficiently informative on its own and that combining it with tabular data introduces noise rather than value. In low-volatility periods, all strategies converge to similar performance levels (Sharpe ratios between 0.29 and 0.54) indicating that calm markets offer limited differentiation between approaches.

Notably, the strategy using FinBERT sentiment shows a consistent improvement as volatility increases (0.30, 0.60, 0.68), suggesting that sentiment data provide additional value with an increase in market stress. This further confirms the important role of sentiment data in predicting aluminum prices, independently of the type of models used. However, more complex models such as the finetuned Qwen model consistently outperform FinBERT, particularly during periods of medium-volatility where the performance gap is most pronounced. In summary, these findings suggest that sentiment improves the predictions of aluminum prices, with variations in its effects depending on the volatility of aluminum prices, the type of news used as a source of sentiment data, and the modeling framework. Therefore, trading strategies can benefit from weighting sentiment inputs based on textual data sources and volatility levels.

4.2. Role of news topics and event types

We further investigate to what extent the topic and type of events can lead to variations in the quality of the predictive signal. We classify each Reuters headline into one of twelve topic categories (*Price Movement, Environmental, Market Analysis, Production Output, Macroeconomic, Inventory Stocks, Demand Outlook, Supply Disruption, Company News, Trade Policy, Geopolitical, and Other*) with the Qwen3 8B base model through zero-shot prompts. We distinguish between predictive statements (forecasts, expectations, guidance) and statements from events that occurred in the past. To assess the contribution of individual news topics to strategy performance, we construct separate portfolios for each topic classification. Specifically, for each month, we filter headlines belonging to a given topic (e.g., *Price Movement*) and compute the monthly sentiment score using only those headlines. This topic-specific sentiment score then determines the trading signal for that month. The resulting monthly returns are used to compute a Sharpe ratio for each topic, allowing us to isolate which types of news carry the most informative sentiment for aluminum price prediction. Months in which no headline of a given topic is available are excluded from that topic's portfolio, which explains the variation in sample sizes across categories.

Figure 3A reports Sharpe ratios for sentiment-only strategies based on the 5 most present individual topics (*Price Movement, Company News, Production Output, Inventory Stocks, Supply Disruption*) compared to the all-topics benchmark. The benchmark strategy, which aggregates sentiment across all headlines, achieves a Sharpe ratio of 0.81. In particular, no individual topic matches this benchmark performance, illustrating the diversification benefits of information aggregation. We extend this analysis by systematically evaluating all possible combinations of 2 to 11 topics to identify an optimal subset of topics, reported in the Figure as "Best Combo". Among 4,094 combinations tested, the best-performing subset includes eight topics: *Company News*, *Supply Disruption*, *Inventory Stocks*, *Price Movement*, *Demand Outlook*, *Geopolitical*, *Macroeconomic*, and *Other*. This combination achieves a Sharpe ratio of 1.00, representing a 23.6% improvement over the all-topics benchmark. The four excluded topics—*Production Output*, *Market Analysis*, *Trade Policy*, and *Environmental*—appear to introduce noise that dilutes signal quality.

Among individual topics, *Price Movement* headlines generate the highest Sharpe ratio (0.71) (Figure 3A), which is intuitive given that such headlines directly concern price dynamics. However, the substantial gap between this topic-specific strategy and the benchmark (0.71 vs. 0.81) indicates that price commentary alone cannot account for potentially valuable signals contained in other news categories. Interestingly, *Supply Disruption* headlines produce a negative Sharpe ratio (-0.07), suggesting that naive sentiment interpretation of disruption news can generate misleading signals, possibly because markets rapidly price in supply-side information or because the sentiment direction does not straightforwardly map to price implications.

Strategies based on forward-looking headlines achieve a nearly zero Sharpe ratio (-0.01), while those based on reports of events that occurred generate a Sharpe ratio of 0.62 (Figure 3B). Forward-looking statements—analyst forecasts, company guidance, demand projections—can reflect expectations that are likely already incorporated into market prices. By contrast, reports of actual events

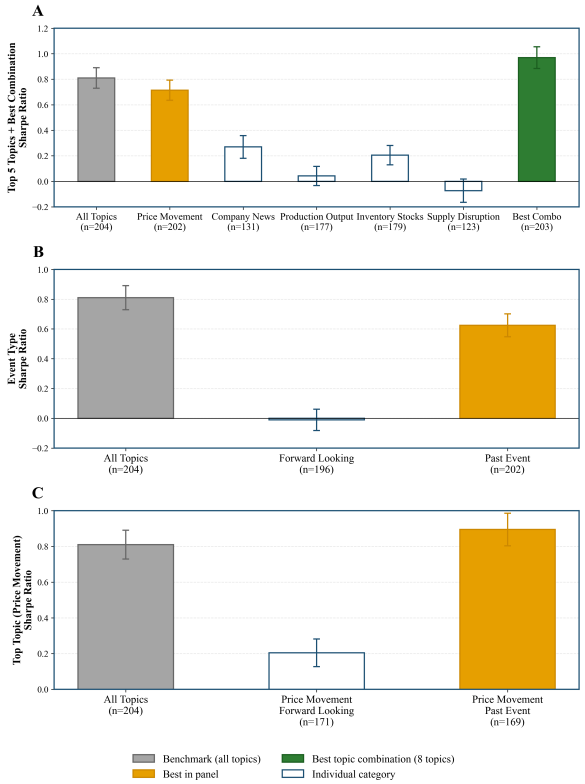


Figure 3: **Evaluation of topic and event type in predictive aluminum prices. A** Comparison of Sharpe ratio across the top five most covered topics (Price Movement, Company News, Production Output, Inventory Stocks, Supply Disruption) and the best topic combination, all calculated using Reuters headlines with the finetuned Qwen sentiment. The benchmark (gray) represents all headlines aggregated. Here n shows the number of months in which each topic is present. **B** Comparison of forward-looking versus past event news types. **C** Comparison of the top-performing topic (Price Movement) by event type. Error bars represent ±1 standard error. The green bar indicates the best-performing topic combination, while the orange bars indicate the best performing individual topic or event type.

(production figures, inventory releases, supply disruptions) can represent new information that cannot be fully anticipated. The near-zero Sharpe ratio for forward-looking content is consistent with the Efficient Market Hypothesis, suggesting that publicly available expectations are already reflected in aluminum prices (Fama, 1970; Tetlock, 2007). We further decompose the dominant *Price Movement* topic (Figure 3C). Within this category, forward-looking headlines yield a Sharpe ratio of 0.20, while occurred events achieve 0.89. This within-topic comparison reinforces the broader finding: actual outcomes carry substantially more predictive content than expectations or forecasts, even when controlling for topic.

In general, these findings suggest that sentiment strategies can benefit from filtering that emphasizes factual reporting over forward-looking commentary. This filtering could be implemented through the event-type classification framework developed in this study, allowing traders to construct signals that prioritize information content over signal volume.

4.3. Role of the source of news

This section evaluates the performance of models using different news sources. Sources can vary in their coverage of topics with high predictive content, and hence lead to varying predictive performance. We compare the predictive performance of Reuters, Dow Jones and China News Service using a sentiment-only strategy with sentiment classified from aluminum related headlines by the same finetuned Qwen3 8B model. Overall, a portfolio using a sentiment-only strategy with Reuters headlines achieved a Sharpe ratio of 0.80 and 433% Cumulative Returns across the whole time period. In contrast, portfolios based solely on sentiment extracted by the same model from headlines of Dow Jones and China News Service achieved a Sharpe ratio of 0.18 and 0.15 and Cumulative Returns of 32% and 22%, respectively.

Figure 4A reports the topic-level distribution of headlines for each source, while Figure 4B shows the global Sharpe ratios aggregated across all sources by topic. The results reveal that the topics with higher ability to predict aluminum price show similarity among news sources. *Price Movement* seems to be the most informative category, achieving a Sharpe ratio of 0.68 and substantially outperforming all other topics, consistent with the intuition that news explicitly referencing price dynamics conveys the most direct signals for future commodity returns. *Environmental* news is second (with a Sharpe ratio of 0.43), indicating that regulatory developments, climate events, and sustainability-related information contain economically meaningful signals, although it being underrepresented in overall coverage (average of 2.3%). In contrast, *Geopolitical* and *Trade Policy* topics exhibit negative Sharpe ratios (-0.26 and -0.25), suggesting limited or counterproductive predictive value, potentially reflecting rapid information diffusion or anticipatory pricing effects.

The differences in overall performance between news sources closely mirror their topic coverage strategies. Reuters allocates a substantial share of its coverage to *Price Movement* (35.6%), the most predictive topic, while maintaining relatively balanced exposure in the remaining categories. By contrast, Dow Jones concentrates nearly half of its coverage (48.0%) on *Company News*, the largest

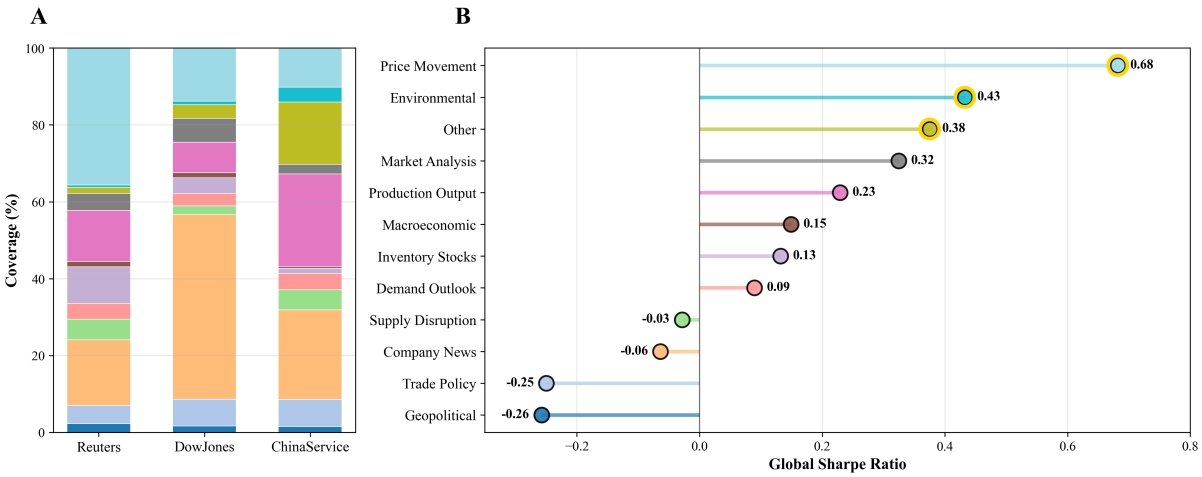


Figure 4: **Predictive performance by topic coverage. A** Percentage of headlines allocated to each topic by source. **B** Global Sharpe ratio by topic, with gold rings highlighting the top three performers. Topics are sorted by global Sharpe ratio (descending).

single-topic allocation among all sources, despite this category showing a negative global Sharpe ratio (-0.06). Added to a limited exposure to *Price Movement* (13.9%), this results in lower predictive performance. ChinaService follows a distinct strategy, emphasizing *Production Output* (24.0%) and *Company News* (23.3%), reflecting its focus on industrial and corporate developments; while *Production Output* exhibits moderate predictive ability (Sharpe ratio of 0.23), a large allocation of negatively performing *Company News* reduces the overall effectiveness of sentiment-based signals.

Although allocating coverage to highly predictive topics is necessary for strong performance, it is not sufficient on its own. Table 1 shows the difference in the Sharpe ratio between sources on the same topics, suggesting that the quality of the information embedded in the news varies substantially between providers. For instance, within the *Price Movement* category, the single most predictive topic overall, Reuters achieves a Sharpe ratio of 0.71, compared to 0.53 for DowJones and 0.67 for ChinaService. This represents a 26% performance gap between Reuters and DowJones on identical topic exposure, highlighting that differences arise not only from topic selection, but also from how information is conveyed. Similar patterns emerge across most topics, where Reuters consistently outperforms competitors despite comparable thematic coverage. These results suggest that the advantage of Reuters might stem from higher signal-to-noise content, clearer causal structure, and more market-relevant timing, enabling sentiment signals to translate more effectively into returns. In contrast, the weaker performance of other sources on the same topics indicates dilution through descriptive, delayed, or less economically grounded reporting.

Topic	Reuters	DowJones	ChinaService	Improvement vs Reuters	
	(Benchmark)	Sharpe	Sharpe	DowJones (%)	ChinaService (%)
Price Movement	0.714	0.525	0.674	-26.6	-5.6
Environmental	0.601	0.347	0.395	-42.3	-34.3
Other	0.612	0.123	0.167	-79.9	-72.8
Market Analysis	0.221	0.348	0.020	+57.5	-91.1
Production Output	0.042	0.125	0.283	+198.5	+577.4
Macroeconomic	-0.051	0.010	0.164	+120.1	+419.8
Inventory Stocks	0.204	0.027	0.222	-86.9	+8.8
Demand Outlook	0.216	-0.258	0.046	-219.5	-78.9
Supply Disruption	-0.073	0.164	0.124	+323.9	+269.6
Company News	0.270	-0.338	0.048	-225.3	-82.3
Trade Policy	-0.391	-0.198	-0.262	+49.3	+33.0
Geopolitical	0.338	-0.387	-0.392	-214.6	-216.1

Table 1: Sharpe ratio comparison across all individual topics, sorted by global performance. Improvements are computed relative to Reuters as the benchmark.

5. Conclusion

We explore the integration of sentiment signals derived from finetuned large language models (LLMs) into the prediction of aluminum prices. Although traditional time-series models that rely exclusively on tabular data provide a solid baseline for market forecasting, our results indicate that incorporating textual sentiment can enhance both predictive accuracy and economic utility, particularly under volatile market conditions. We introduce comprehensive sentiment based price forecasting of aluminum, a commodity that typically has much lower trade volume and media coverage than other metals such as gold or silver.

The finetuned Qwen3 8B consistently generates trading signals that outperform FinBERT in our out-of-sample portfolio simulations. This aligns with better sentiment classification results by finetuned LLMs compared to FinBERT reported by the literature and suggests that this is translated into improved risk-adjusted returns. Sentiment signals provided mixed value during periods of different volatility. In medium volatility regimes, signals de-

rived from sentiment alone outperformed time series forecasting models as well as those combined with sentiment. This suggests that in certain market conditions, sentiment alone can be enough as a price directionality predictor. However, in high volatility regimes, while LSTM models using tabular data only exhibited diminished performance, when sentiment data is added, it achieves the best Sharpe ratio. These results indicate that sentiment can provide complementary information that is unlikely to be captured by numerical time-series patterns alone during volatile periods.

This study also explores the importance of the source of sentiment. Differences in trading performance across Reuters, Dow Jones, and China News Service are explained not only by topic allocation, but also by information quality. Reuters consistently delivers higher Sharpe ratios even when controlling for topic exposure, suggesting superior signal-to-noise characteristics and more relevant framing. This finding suggests that sentiment modeling performance depends jointly on language model quality and upstream information selection. In addition, filtering news by topic can substantially improve strategy performance, as demonstrated by a 23.6% Sharpe improvement over the all-topics benchmark when excluding certain topics.

However, we acknowledge that the sparsity of coverage of aluminum related events by news sources limits the temporal granularity of this study to monthly frequency. Future research could extend this framework to other metals or commodities as well as integrate higher frequency sentiment streams to capture finer temporal dynamics. Overall, our results provide evidence that carefully calibrated sentiment analysis using finetuned LLMs can serve as a complementary and in some cases the main source of predictive information in commodity markets, particularly under conditions where traditional numerical indicators are less informative.

6. Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant Nos. T2350610281 and 82273731).

7. Bibliographical References

Nichole Bastin. 2024. Aluminum mmi: Aluminum prices rise, downside risks remain.

Svetlana Borovkova. 2011. News analytics for energy futures. *SSRN Electronic Journal*.

Christian Breitung, Garvin Kruthof, and Sebastian Müller. 2023. Contextualized sentiment analysis using large language models. *SSRN Electronic Journal*.

Y. Chen, H. Zhang, and Y. Li. 2020. Analyzing the impact of public sentiment on stock performance: Evidence from china's stock market. *Finance Research Letters*, 34:101254.

Y. Chen, J. Zhang, and X. Wang. 2021. Sentiment analysis for copper price forecasting based on social media data. *Resources Policy*, 70:102946.

J. M. Cullen and J. M. Allwood. 2013. Mapping the global flow of aluminum: From liquid aluminum to end-use goods. *Environmental Science & Technology*, 47(7):3057–3064.

Stavros Degiannakis, Peter Dent, and Christos Floros. 2014. A monte carlo simulation approach to forecasting multi-period value-at-risk and expected shortfall using the figarch-skt specification. *The Manchester School*, 82(1):71–102.

Gillian Dooley and Helena Lenihan. 2005. An assessment of time series methods in metal price forecasting. *Resources Policy*, 30(3):208–217.

Moses Olabhele Esangbedo, Blessing Olamide Taiwo, Hawraa H. Abbas, Shahab Hosseini, Mohammed Sazid, and Yewuhalashet Fissha. 2024. Enhancing the exploitation of natural resources for green energy: An application of lstm-based meta-model for aluminum prices forecasting. *Resources Policy*, 92:105014.

Christian-Oliver Ewald and Yifan Li. 2024. The role of news sentiment in salmon price prediction using deep learning models. *Journal of Commodity Markets*, 36:100438.

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

S. Fatemi, Y. Hu, and M. Mousavi. 2024. A comparative analysis of instruction fine-tuning llms for financial text classification. *arXiv preprint arXiv:2411.02476*.

R. Gupta, S. Kumar, and A. Singh. 2020. Impact of social media sentiments on gold prices: An empirical investigation using twitter data. *Journal of Economic Behavior & Organization*, 179:108–123.

James D. Hamilton. 1994. *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Gao Huang. 2024. Dynamic neural networks: Advantages and challenges. *National Science Review*, 11(8):nwae088.

J. Inger, M. Kumar, and A. Somani. 2018. Optimal historical window length in commodity futures return forecasting. *Journal of Commodity Markets*, 11:1–16.

International Aluminium Institute. 2021. Beyond 2 degrees: The outlook for the aluminium sector factsheet.

G. Kant, I. Zhelyazkov, A. Thielmann, C. Weisser, M. Schlee, C. Ehrling, B. S"afken, and T. Kneib. 2024. One-way ticket to the moon? an nlp-based insight on the phenomenon of small-scale neobroker trading. *Social Network Analysis and Mining*, 14(1):121.

Thanos Konstantinidis, Giorgos Iacovides, Mingxue Xu, Tony G. Constantinides, and Danilo Mandic. 2024. Finllama: Financial sentiment classification for algorithmic trading applications. *arXiv preprint arXiv:2403.12285*.

Thomas Kriechbaumer, Andrew Angus, David Parsons, and Monica Rivas Casado. 2024. An improved wavelet-arima approach for forecasting metal prices. *Journal of Forecasting*, 44(2):253–270. Corresponding author: m.rivas-casado@cranfield.ac.uk.

A. Kumar and V. Ravi. 2021. Stock price prediction using financial news sentiment analysis: A machine learning approach. *Expert Systems with Applications*, 165:113845.

S. Kumar and S. Singh. 2024. Fine-tuning llama 3 for sentiment analysis: Leveraging aws cloud for enhanced performance. *SN Computer Science*, 5.

Y. Lan et al. 2023. Chinese fine-grained financial sentiment analysis with large language models. *arXiv preprint arXiv:2306.14096*.

Zixuan Li, Yu Yang, Yue Chen, and Jun Huang. 2023. A novel non-ferrous metals price forecast model based on lstm and multivariate mode decomposition. *Axioms*, 12:670.

Andrew W. Lo. 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5):15–29.

G. M. E. Luglio. 2023. The global price of aluminum: A dynamic journey over the past few years.

Stina Johanne Mysen and Elisabeth Marie Thornton. 2021. Forecasting the price of aluminium using machine learning: Empirical comparison of machine learning and statistical methods.

K. Oikonomou and D. Damigos. 2024. Short term forecasting of base metals prices using a lightgbm and a lightgbm-arima ensemble. *Mineral Economics*.

Alvaro Paredes Amorin, Andre Python, and Christoph Weisser. 2025. Fine-tuning of lightweight large language models for sentiment classification on heterogeneous financial textual data. *arXiv preprint*. Preprint.

Robert S. Pindyck. 1999. The long-run evolution of energy prices. *The Energy Journal*, 20(2):1–27.

N. S. Sinatrya, I. Budi, and A. Budi Santoso. 2022. Classification of stock price movement with sentiment analysis and commodity price: Case study of metals and mining sector. In *2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 59–64.

Harald U. Sverdrup, Kristin Vala Ragnarsdottir, and Deniz Koca. 2015. Aluminium for the future: Modelling the global production, market supply, demand, price and long term development of the global reserves. *Resources, Conservation and Recycling*, 103:139–154.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

M.-L. Thormann, J. Farchmin, C. Weisser, R.-M. Kruse, B. S"afken, and A. Silbersdorff. 2021. Stock price predictions with lstm neural networks and twitter sentiment. *Statistics, Optimization & Information Computing*, 9(2):268–287.

N. Wang, H. Yang, and C. D. Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Xiaojie Yi, Yonglong Lu, and Guizhen He. 2024. Aluminum demand and low carbon development scenarios for major countries by 2050. *Journal of Cleaner Production*, 475:143647.

F. C. Yuan, C. H. Lee, and C. Chiu. 2020. Using market sentiment analysis and genetic algorithm-based least squares support vector regression to predict gold prices. *International Journal of Computational Intelligence Systems*, 13:234–246.

B. Zhang, H. Yang, and X.-Y. Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

B. Zhang et al. 2023b. Enhancing financial sentiment analysis via retrieval augmented large language models. *arXiv preprint arXiv:2310.04027*.

Appendix A: Filtering aluminum news data with LLaMA

News on aluminum are directly collected from Factiva from Dow Jones Newswires and Reuters. However, since it is a bulk retrieval of 7000-10000 news per source, there are many entries that do not contain any relevant information and can impact the sentiment score results. Therefore, the two aluminum news datasets are filtered using the LLM LLaMA3 8B in a few-shot scenario.

The figure below shows the UMAP-HDBSCAN Clustering of both datasets before and after applying the filter. The clustering figure shows how after applying this filter, outliers are reduced substantially. Below each graphic, some examples from 5 different cluster groups can be seen. Sentences in the same group have high similarity, as can be seen in the examples below. In the filtered datasets, the number of clusters is reduced because some of the clusters are removed from the dataset. However, the method used with LLaMA3 does not discern between clusters but only the overall meaning of the sentence, and it is highly influenced by the examples given in the few-shot prompt. Therefore, combined with professional expertise, this method can be highly efficient in filtering large amounts of data in a fraction of time compared to doing it manually or with word mapping strategies.

Figure 5: Cluster plot of the Reuters and Dow Jones Newswires aluminum datasets before (left) and after (right) filtering using LLaMA3. The y and x axis are the 2 dimensions obtained by reducing the embeddings dimensions. The plots on the left show less number of outliers outside the main groups clusters.

Appendix B: Time Series Models Training and Testing

The training methodology begins with data preprocessing, where the dataset containing metal prices and economic indicators is loaded, sorted by date, and missing values are imputed. The dataset is then resampled to generate monthly time series, adding sentiment score variables from each of the aluminum news sources (Reuters, DowJones Newswires and China News Service) and leaving one baseline dataset without sentiment variable. Rolling windows of 4 different lengths (1-month, 3-months, 6-months, and 12-months) are created to generate sequences of input features and corresponding target values, reshaped to match the input requirements of the models. This, added to the 5 different models and 4 sentiment sources, leaves us with 80 combinations, represented in Figure 6. The models are trained using mean squared error loss and the Adam optimizer, with gradient clipping and learning rate scheduling applied to ensure stable convergence.

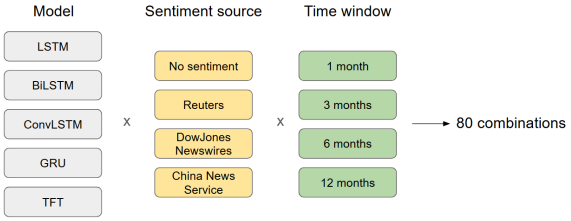


Figure 6: Combination of models, sentiment sources and time windows

The evaluation methodology uses walk-forward validation, a technique designed specifically for time series forecasting. In this approach, the model is trained on a rolling window of past observations, starting with an initial segment of the historical data. After training in this window, the model predicts the value immediately following the training period. The window then slides forward to include the new data point, and the process repeats iteratively for each subsequent time step. This method ensures that at each prediction, the model only has access to information that would have been available at that time, closely mimicking real-world forecasting scenarios and preventing data leakage, as well as maximizing the testing period.

A grid search or hyperparameter search is conducted over hidden sizes, numbers of layers, and dropout values, with results recorded for each configuration. The chosen ranges are 16, 32, 64, 128 and 256 for hidden size and 1 to 6 for the number of layers. The dropout value used is always 0.1. Therefore, added to the original 80 setup combinations makes up a total of 2400 combinations of

variables and parameters. However, only the best result for every hidden size and number of layers combination is collected, leaving 80 best results in total. The best-performing model for each window size is retrained on the full dataset and stored along with its predictions and evaluation metrics. This approach ensures robust model assessment, accounts for temporal dependencies, and minimizes data leakage in time series forecasting tasks.

The models performance are assessed using metrics such as R^2, RMSE and MAE. These are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{6}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7}$$

where y_i denotes the true observed value for index i, $\hat{y}_i$ is the prediction of the corresponding model, $\bar{y}$ represents the mean of the observed values, i indexes each observation and n is the total number of observations.

Appendix C: Time Series Models Results

C.1. Results for hidden size and number of layers from the grid search

The best result for every combination of hidden size and number of layers is reported in Appendix E in Appendix B, reporting the best combination of hyperparameters.

Regarding hidden size and number of layers, we find that the average optimal values are around 3 layers and 128 hidden size, varying between models. These are shown in Figure 7. A bigger number of layers tends to overfit, as well as hidden size values exceeding 128.

C.2. Aluminum Price Forecasting Results

Since the walk-forward validation method allows us to test with almost the whole dataset, we also gather the predictions from each best model, sentiment source, and time window. Figure 8 represents the observed aluminum price (solid line) and the aluminum price predicted by LSTM (dashed line) within a time window of 6-months and using sentiment scores from Reuters news.

To study the impact of the sentiment variable and different time windows, we averaged the results

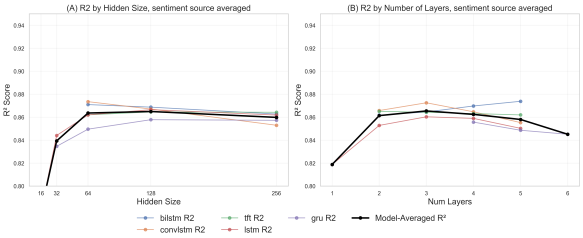


Figure 7: Average R^2 score by number of layers and by hidden size. The solid line in black is the average of the five time series models. The faded lines are each respective model R^2, blue for BiLSTM, green for TFT, purple for GRU, orange for ConvLSTM and red for LSTM.

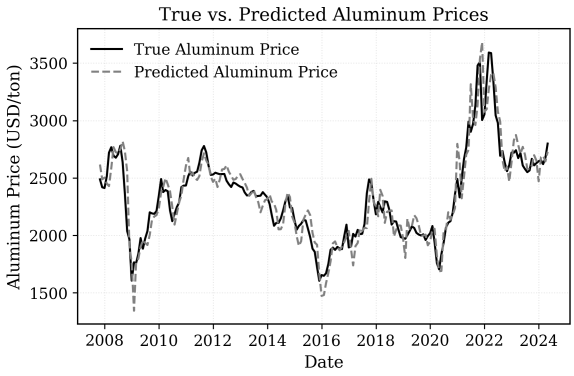


Figure 8: Predicted aluminum price and true aluminum price from November 2007 to April 2024.

by sentiment source and time window per model. These results are represented in Fig 9.

The observed high predictive performance of models using short historical windows for next-month forecasting aligns with established principles in time series econometrics. Commodity prices exhibit time-varying autocorrelation structures where recent observations contain the strongest predictive signals, while distant historical data often introduce noise rather than information, a phenomenon documented in financial time series analysis ((Hamilton, 1994)). This pattern is particularly pronounced in metal markets, where short-term dynamics is driven by recent supply-demand shocks and sentiment, while longer historical patterns reflect structural breaks and regime changes that provide limited incremental predictive power for immediate forecasts ((Pindyck, 1999)).

The preference for recent data over extended historical windows reflects optimal feature selection in non-stationary financial environments. As established in commodity forecasting research ((Borovkova, 2011)), financial markets mix persistent fundamentals with transient noise, where short windows effectively isolate relevant signals while

discarding historical noise that could dilute predictive accuracy. The observed improved performance of 6-month windows on both shorter (3-month) and longer (12-month) alternatives aligns with previous findings ((Degiannakis et al., 2014)), who demonstrate that intermediate historical windows optimally balance short-term noise filtering with sufficient context to capture business cycle patterns in metal markets. This intermediate-length window provides enough data to identify meaningful cyclical patterns while avoiding regime changes and structural breaks that increasingly contaminate longer historical series. This approach also mitigates overfitting to spurious patterns that appear significant in the sample but do not generalize, a critical concern in financial forecasting identified in empirical market studies ((Lo, 2004)). The methodology aligns with evidence showing diminishing marginal information gains from additional historical data beyond intermediate windows ((Inger et al., 2018)), creating an effective trade-off between model complexity and forecasting robustness where 6-month windows can represent a sweet spot for aluminum price prediction.

The aggregated results reveal clear performance hierarchies and interaction effects between model architectures and sentiment sources. ConvLSTM emerges as the architecture that performs the best overall, achieving the highest average R^2 of 0.8902 when paired with ChinaService sentiment data, while also providing the strongest accuracy for monthly predictions (0.9022) when averaged across all sentiment sources. BiLSTM demonstrates remarkable consistency, maintaining nearly uniform performance across time windows (0.8788-0.8888) and excelling particularly with no-sentiment data (0.8870 average). GRU exhibits the greatest sensitivity to external information, showing a substantial 0.0538 performance gap between sentiment-enhanced configurations and the no-sentiment baseline, indicating strong dependency on external feature engineering. Interestingly, the no-sentiment baseline proves surprisingly competitive across multiple architectures, matching or even exceeding sentiment-augmented performance for LSTM, BiLSTM and ConvLSTM, suggesting that raw price history contains sufficient predictive signals for these models, while sentiment features can introduce conflicting noise rather than pure signal enhancement.

The analysis further uncovers distinct temporal performance patterns and source-model synergies. One-month prediction windows consistently yield the best results across architectures, supporting the temporal locality hypothesis, where recent data contains the strongest predictive signals. Reuters sentiment emerges as the most reliable information source, providing consistently strong performance across all models with minimal variability, while ChinaService shows specific synergy with ConvLSTM architecture. Performance declines with 3 months historical windows and grows back with 6 months windows.

C.3. Portfolio results

We apply the price-based trading strategy described in section 3.3 to the monthly aluminum price predictions made by the best hyperparameter combination found during the grid search for each type of model, sentiment source, and time window, gathered in Appendix E. Then we estimated R^2, RMSE, MAE, Hit Rate, p value, and total returns over the period of time in Appendix F. These metrics will be discussed in the following section.

C.3.1. Hit rate

Hit rate analysis provides a direct measure of a model's predictive quality, independent of position sizing, transaction costs, and market volatility effects. The hit rate refers to the percentage of forecasts in which the model correctly predicts the direction of price movement (up or down) over a given horizon. A consistently high hit rate (>0.50) demonstrates that the model captures meaningful directional signals beyond random chance, offering a systematic edge crucial for long-term trading viability. While profitability metrics like total returns and Sharpe ratios evaluate overall performance, hit rates reveal whether the strategy's success stems from genuine predictive power or merely from risk management and occasional large wins. This distinction is vital for strategy robustness, as models with low hit rates but positive returns often depend on unsustainable market conditions or excessive risk-taking, whereas high hit rates indicate reliable signal generation that can be scaled and optimized with proper execution.

The hit rate results reported in Table 2 reveal a systematic hierarchy in predictive accuracy across model architectures and sentiment sources. The 3-month forecasting window consistently delivers peak performance across both models and sentiment sources, with Reuters sentiment achieving the highest overall hit rate of 0.577 at this horizon. Among architectures, ConvLSTM shows a better short-term forecasting capability with a 1-month hit rate of 0.585 and the highest model-average accuracy (0.564), while BiLSTM performs best at 3-months horizons. Notably, the no-sentiment baseline maintains competitive hit rates across all time windows and achieves a source-average of 0.551, only slightly below sentiment enhanced sources, suggesting that historical price patterns contain substantial directional information independent of sentiment signals. The observed stability in the hit rates

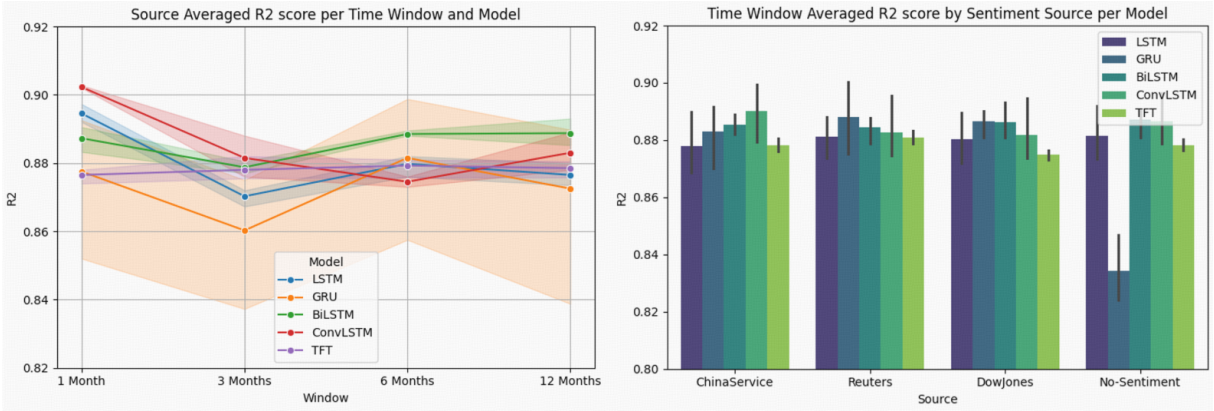


Figure 9: Sentiment source averaged R^2 score per time window and model on the left, bands show the 95% confidence intervals of those mean R^2 values. Time window averaged R^2 score per sentiment source and model on the right, whiskers show the 95% confidence intervals of those mean R^2 values.

on different horizons and configurations indicates robust predictive capabilities, with all configurations maintaining statistically significant accuracy above random chance and providing an overall average hit rate of 0.556 (55.6%) regardless of window length or data source.

Configuration		Time Window				
Category	**Name**	**1M**	**3M**	**6M**	**12M**	**Avg**
A. Model Performance						
Models	BiLSTM	0.549	0.573	0.562	0.534	0.554
	ConvLSTM	0.585	0.569	0.559	0.543	**0.564**
	GRU	0.550	0.545	0.562	0.545	0.551
	LSTM	0.564	0.563	0.549	0.548	0.556
	TFT	0.554	0.568	0.538	0.560	0.555
	Model Avg	0.560	**0.563**	0.554	0.546	0.556
B. Sentiment Source Performance						
Sources	ChinaSvc	0.570	0.567	0.558	0.550	**0.561**
	DowJones	0.551	0.549	0.555	0.554	0.552
	No-Sent.	0.564	0.561	0.543	0.536	0.551
	Reuters	0.557	0.577	0.560	0.543	0.559
	Source Avg	0.560	**0.563**	0.554	0.546	0.556

Table 2: Directional Hit Rate Performance by Model and Sentiment Source Across Time Windows. Hit rates represent the percentage of correct directional predictions. Values closer to 1.000 indicate better performance. The 3-month window shows the highest average hit rate across both models and sources. Values in bold highlight the time window (panels A and B) with the best average hit rate, and the model (panel A) and sentiment source (B) with the best average hit rate.

C.3.2. Average returns

The return performance analysis presented in Table C.3.2 reveals substantial heterogeneity in both the forecasting efficacy and the strategic value of the models evaluated. In particular, these results represent the average performance across all tested configurations rather than the maximum achievable performance under ideal settings. Consequently, many of the reported return metrics are negative or near zero, reflecting the dilution effect of including underperforming setups in the ensemble average. In Table C.3.2A, the results shown are the sentiment averaged returns for each model and time window. LSTM performs best in the 1-month horizon, achieving a 168.7% return (1.687 multiplicative factor), while GRU shows the best performance at 6-months time window with a 108.7% return (1.087). GRU exhibits the most volatile performance profile, transitioning from -30.8% at 3 months to 108.7% at 6 months. TFT consistently underperforms on all time horizons, yielding predominantly negative returns and the lowest average performance of -42.9% (-0.429). BiLSTM and ConvLSTM show mixed results, with BiLSTM achieving moderate success at shorter horizons (52.3% at 3 months) but declining at longer horizons, while ConvLSTM shows minimal positive performance only at the 12-month horizon (14.7%). The best performing models can reach high returns: LSTM (292%), GRU (266%) and BiLSTM (232%). Despite their relatively high complexity, TFT and ConvLSTM tend to underperform, indicating that higher model complexity does not guarantee higher returns.

In Table C.3.2B, the model averaged returns for each sentiment source and time window reveal that sentiment enhanced models generally outperform the no-sentiment baseline. Reuters sentiment emerges as the most effective source overall, achieving a 28.2% average return and showing particularly strong medium-term performance with 55.7% return at the 6-month horizon. The sentiment of ChinaService demonstrates strong short-term efficacy with 58.8% return at 1 month, but experiences performance deterioration over longer horizons. In particular, all sentiment sources, in-

cluding the baseline without sentiment, exhibit negative or almost negative performance at the 12-month horizon, with Dow Jones sentiment showing the most significant decline (-45.9% return). This consistent pattern suggests that while sentiment integration can enhance short-to-medium term forecasting, its predictive value diminishes over longer horizons. The results, visualized in Figure 10, show how the sentiment of Reuters peaks within a medium term window, while predictive accuracy remains systematically poor over longer time windows.

Configuration		Time Window					
Category	Name	1M	3M	6M	12M	Avg	Best
A. Model Performance							
Models	BiLSTM	0.346	0.523	0.097	-0.076	0.222	2.32
	ConvLSTM	-0.292	0.062	-0.096	0.147	-0.045	1.02
	GRU	0.125	-0.308	1.087	-0.334	0.142	2.66
	LSTM	**1.687**	-0.178	0.105	-0.269	**0.336**	**2.92**
	TFT	-0.239	-0.559	-0.412	-0.504	-0.429	0.05
	Avg	**0.325**	-0.092	0.156	-0.207	0.045	
B. Sentiment Source Performance							
Sources	ChinaSvc	**0.588**	-0.299	0.085	0.067	0.110	1.52
	DowJones	0.148	-0.171	0.003	-0.459	-0.120	1.36
	No-Sent.	-0.021	-0.233	-0.020	-0.085	-0.090	1.31
	Reuters	0.587	0.334	0.557	-0.352	**0.282**	**2.92**
	Avg	**0.325**	-0.092	0.156	-0.207	0.045	

Table 3: Total return performance by model and sentiment source across time windows. **A.** Sentiment-source averaged returns for each mode. **B.** Model-averaged returns for each sentiment source. Values represent total returns (multiplicative factors, where a value of X indicates an X$\times$100% return). The "Avg" column shows time-window averaged performance, and "Best" reports the highest return achieved.

The relationship between directional accuracy (hit rate) and realized returns reveals a complex association between predictive quality, given by the hit rate, and financial performance, given by the returns. ConvLSTM, although it delivers the best average hit rate, achieves the second lowest average returns (-4.5%). This divergence suggests that while hit rates provide fundamental predictive validation, and therefore all models maintain statistically significant accuracy above random, financial performance requires additional mechanisms beyond directional correctness. The 1-month forecast window consistently optimizes both metrics, achieving the second best average hit rate (0.560) and highest average return (32.5%), indicating that the latest signals are most important when predicting future prices. Sentiment addition similarly enhances both dimensions, with all sentiment sources delivering a higher average hit rate, higher returns, and higher best returns. In particular, the best performing setups are the ones that add sentiment scores from Reuters.

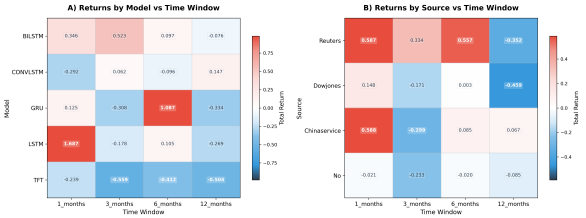


Figure 10: Model averaged returns by time window (A). These are the total returns of all setups averaged for each model and time window. Source averaged returns (B) by time window. These are the total returns of all setups averaged for each sentiment source and time window.

C.3.3. Best portfolio performance from no-sentiment vs sentiment leveraged model

While the previous results show how adding sentiment from either Reuters, Dow Jones or China News Service improves the model predictions R^2, hit rate and total returns across the whole time interval of study (2007 - 2024) when averaged by sentiment source or model, this section compares the results of the configuration of hyperparameters, model type and time window, which gives the best total returns without using sentiment scores and when sentiment scores from a source (Reuters, Dow Jones or China News Service) are used.

Without the use of sentiment scores, which we refer to as "no-sentiment" configuration, the best total returns were achieved with the predictions made by the LSTM model in a 1-month time window, 128 hidden size and 4 layers. This model achieved a R^2 score of 0.90, RMSE of 120.17, and MAE of 84.28. After applying the same simulation of the trading strategy as described in 3.3 to its predicted monthly aluminum prices, it achieved a total return of 131%.

However, when including sentiment, the highest total returns were achieved by the configuration of the LSTM model in a 1-month time window, 128 hidden size, and 4 layers with sentiment from Reuters. Initially, it got an R^2 score of 0.89, RMSE of 125.12, and MAE of 87.98, which is slightly worse than the no-sentiment variant, whereas after running the trading simulation the total returns outperform, achieving a total return of 292%.

The substantial increase in total returns (292% vs 131%) observed when incorporating Reuters sentiment, despite marginally lower R^2, RMSE and MAE scores, can be attributed to the directional accuracy and timing precision of the model's forecasts, which are not fully captured by point-prediction error metrics.

R^2, RMSE and MAE measure the magnitude of prediction errors in all time steps, penalizing devia-

tions in predicted price levels regardless of whether the forecast correctly anticipates the direction of price movement. In contrast, trading performance depends critically on correctly predicting price direction at key turning points—particularly around market reversals, news events, or sentiment shifts.

Although the no-sentiment model can achieve slightly better average error metrics, the Reuters-augmented model appears to generate more accurate signals at economically meaningful moments, such as identifying impending price increases or decreases that lead to profitable trading decisions. This suggests that sentiment information helps the model better align its predictions with market-moving events reflected in the news flow, even if it introduces small increases in the average prediction error.

Figure 11 represents both the portfolio performance over time and the evolution of the aluminum price. Although both follow a similar pattern, the configuration that adds Reuters sentiment outperforms due to several different trade decisions. Figure 12 shows the monthly return comparison between both variants. Most of the time, the signals are the same for both configurations (with and without sentiment); however, whenever the configurations disagree, the model with sentiment is more often correct. The red lines represent a signal disagreement, where the sentiment and no-sentiment models execute a different trade. Some of the most determinant trades were executed between the 2020 and 2022 time periods. The three largest returns gaps between both models were in March 2020, January 2022, and can 2022.

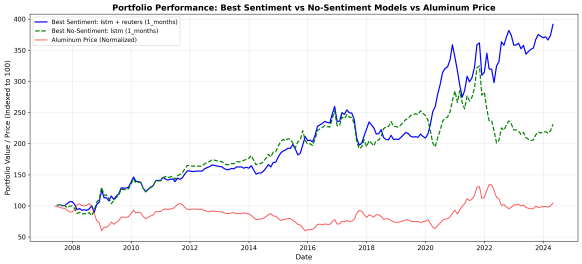


Figure 11: Comparison of the best Sentiment vs No-Sentiment Models returns vs True Aluminum price. Y axis is the Portfolio Value, with the starting point being 100 (100% of the intial value).

To illustrate the results, take month February 2020, with an observed aluminum price of 1940.50$/ton. While the no-sentiment model predicted a 1974.37$/ton price for March and the reuters sentiment model predicted a 1899.57$/ton, with an observed price for March 2020 of 1759.43$/ton. Here the long signal captured by the model without sentiment leads to a -9.33% loss in returns, whereas the short signal captured by the

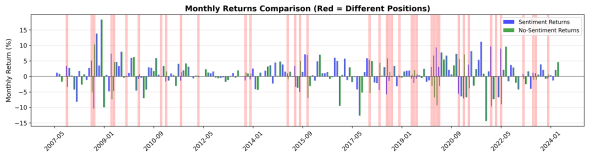


Figure 12: Monthly return of best sentiment and no-sentiment models monthly returns. The red bars show month where the two configurations made a different decision.

sentiment model generated a 9.33% return.

Looking at the sentiment variable that was used to make the predictions, the Reuters sentiment score for February 2020 took the lowest possible value (-1), which is expected to decrease the predicted price value for next month. Table C.3.3 shows the headline, date and associated sentiment generated by the finetuned language model (SFT Qwen3 8B) for February 2020.

News Headline	Date	Sentiment
Coronavirus will negatively affect aluminum market in China	28-Feb-20	Negative
Coronavirus double shock for aluminium sector	24-Feb-20	Negative
LME aluminium can test support at $1,688	24-Feb-20	Negative
Japan aluminium stocks down 2.8%	18-Feb-20	Negative
LME aluminium testing support at $1,709	10-Feb-20	Negative
LME aluminium seeking support at $1,709	03-Feb-20	Negative
Shanghai metals limit-down amid coronavirus fears	03-Feb-20	Negative

Table 4: Sentiment analysis of market news (February 2020).

Let us analyze January 2022 as a second illustration of the results. The model without sentiment model had predicted 3044.21$/ton while sentiment model predicted 3088.49$/ton. The initial price on December 2021 was 3052.88$/ton and the price on January 2022 surged to 3347.41$/ton. Therefore, the no-sentiment model suggested a short trade while the sentiment one suggested a long trade, leading to -9.64% and +9.64% return respectively. The predictions made by the finetuned Qwen3 model, on December of 2021, predicted a sentiment score of 0.18, slightly positive. Table C.3.3 shows all headlines, data, and associated sentiment for the investigated month.

News Headline	Date	Sentiment
Norway's Hydro to cut Slovakia aluminium output further due to power prices	30-Dec-21	Negative
Copper slips in range-bound trade, aluminium shines on supply worries	30-Dec-21	Negative
Power price surge pushes aluminium to 2-month high	23-Dec-21	Positive
China Nov aluminium output at 3.10 mln tonnes – stats bureau	15-Dec-21	Neutral
Scarce supplies to propel aluminium to top LME leaderboard	15-Dec-21	Positive
Marubeni sees Japanese aluminium premiums at $140–$250/T in 2022	07-Dec-21	Neutral
Aluminium prices firm as China plans hiking coal contract prices	03-Dec-21	Positive
Japan aluminium stocks in October up 1.1% m/m – Marubeni	03-Dec-21	Positive
London aluminium edges higher as stockpiles fall, demand recovers	02-Dec-21	Positive
Aluminium dips on Omicron fears, but low inventories cushion fall	02-Dec-21	Negative
Carbon brakes aluminium supply response to booming prices: Andy Home	01-Dec-21	Neutral

Table 5: Sentiment analysis of market news (December 2021).

As a last example, let us consider April 2022,

with an initial price of 3345.02$/ton. While the no-sentiment model predicted a 3374.74$/ton price for can the Reuters-based sentiment model predicted a 3326.42$/ton price. In the end, the real price for March was 3044.82$/ton. This triggered a long signal for the no-sentiment variant, losing -8.97% whereas the sentiment model captured a short signal, earning a positive amount of 8.97%.

On April 20, 2022, the sentiment score was -0.625, which is a relatively strong negative value. Table C.3.3 shows the news headline, date and associated value for that month.

News Headline	Date	Sentiment
London aluminium poised for worst month in over a decade on growth risks	29-Apr-22	Negative
China Shenhuo to raise aluminium output in Yunnan as power curbs ease	26-Apr-22	Positive
Global aluminium output falls 1.55% in March year on year, IAI says	20-Apr-22	Negative
Shanghai aluminium hits 3-month low on strong dollar, demand woes	12-Apr-22	Negative
Shanghai aluminium sinks to 3-month low as demand woes linger	12-Apr-22	Negative
China demand angst hits aluminium prices	11-Apr-22	Negative
Shanghai aluminium slips to over 3-week low as demand concerns weigh	08-Apr-22	Negative
Japan aluminium buyers to pay lower premiums for April–June imports	07-Apr-22	Negative
Japan buyers agree to Q2 aluminium premium of $172/T, sources say	07-Apr-22	Neutral

Table 6: Sentiment analysis of market news (April 2022).

Appendix D: Defined Volatility Regimes Across Time

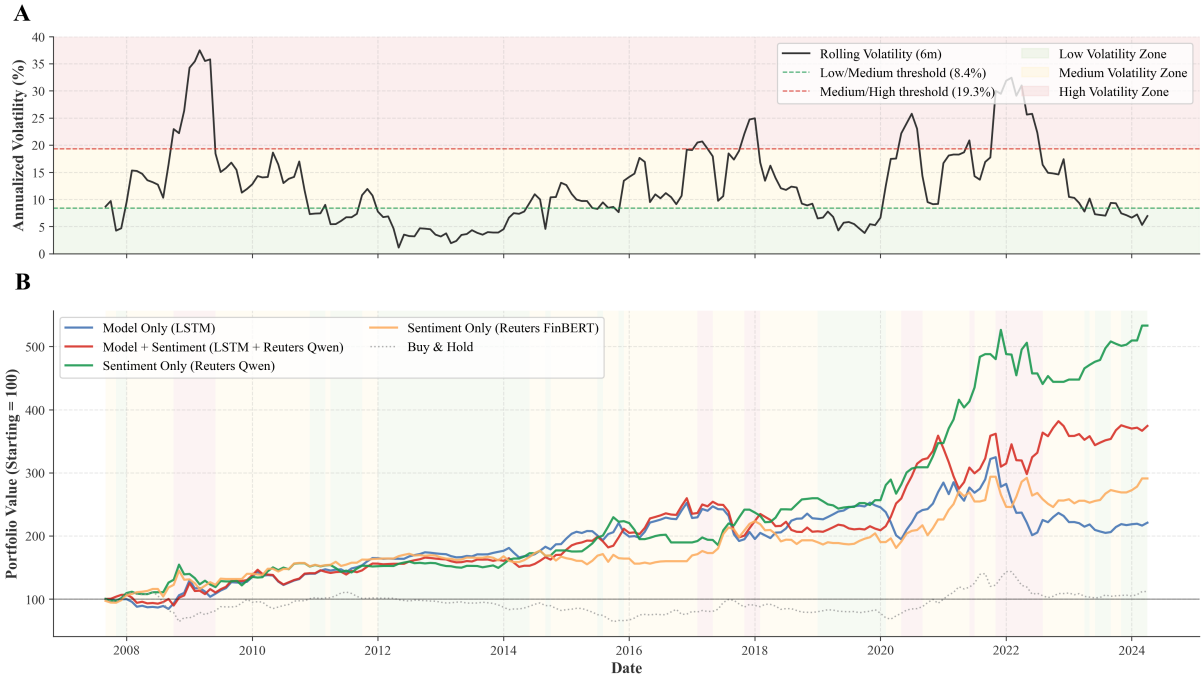


Figure 13: Panel A: Volatility Regimes Over Time. Red area represents the high-volatility regime, yellow represents medium-volatility and green low-volatility. Panel B: Portfolio Values Over Time for the different strategies. Blue line represents the price-based strategy, red line price-based strategy combined with sentiment, green line sentiment only strategy with sentiment predicted by finetuned Qwen, yellow line sentiment only strategy with sentiment predicted by FinBERT and black line buy & hold strategy.

Appendix E: Grid Search best results per model, time window and sentiment source

Model	Source	Horizon	Hidden Size	Layers	R^2	RMSE	MAE
bilstm	reuters	1 months	256	5	0.89	126.79	91.59
bilstm	reuters	3 months	256	3	0.88	132.66	92.02
bilstm	reuters	6 months	64	3	0.89	126.92	88.74
bilstm	reuters	12 months	64	3	0.89	128.64	95.47
bilstm	dowjones	1 months	256	5	0.88	130.12	93.37
bilstm	dowjones	3 months	128	4	0.88	129.93	92.71
bilstm	dowjones	6 months	128	2	0.89	126.94	88.22
bilstm	dowjones	12 months	128	5	0.89	123.87	91.57
bilstm	chinaservice	1 months	64	5	0.89	125.11	89.85
bilstm	chinaservice	3 months	256	5	0.88	130.69	95.01
bilstm	chinaservice	6 months	256	5	0.89	126.80	96.00
bilstm	chinaservice	12 months	128	2	0.88	130.00	98.25
bilstm	no-sentiment	1 months	64	4	0.89	124.68	90.66
bilstm	no-sentiment	3 months	64	3	0.88	131.74	89.92
bilstm	no-sentiment	6 months	128	4	0.89	125.87	89.84
bilstm	no-sentiment	12 months	128	4	0.89	127.39	91.64
convlstm	reuters	1 months	64	3	0.90	117.84	82.64
convlstm	reuters	3 months	128	3	0.88	132.64	90.43
convlstm	reuters	6 months	64	3	0.87	134.80	96.54
convlstm	reuters	12 months	64	3	0.88	132.48	97.26
convlstm	dowjones	1 months	64	3	0.90	117.84	82.64
convlstm	dowjones	3 months	128	3	0.88	132.64	90.43
convlstm	dowjones	6 months	64	3	0.87	134.80	96.54
convlstm	dowjones	12 months	64	3	0.88	134.74	97.52
convlstm	chinaservice	1 months	64	3	0.90	117.84	82.64
convlstm	chinaservice	3 months	64	3	0.89	124.18	87.29
convlstm	chinaservice	6 months	128	5	0.87	134.19	97.63
convlstm	chinaservice	12 months	64	3	0.89	125.63	93.95
convlstm	no-sentiment	1 months	64	3	0.90	117.59	82.30
convlstm	no-sentiment	3 months	128	3	0.88	129.37	89.36

Continues on next page

Model	Source	Horizon	Hidden Size	Layers	R^2	RMSE	MAE
convlstm	no-sentiment	6 months	64	4	0.88	132.81	96.95
convlstm	no-sentiment	12 months	128	5	0.88	130.05	96.46
gru	reuters	1 months	64	6	0.89	126.34	89.13
gru	reuters	3 months	64	4	0.87	135.72	98.21
gru	reuters	6 months	256	4	0.90	117.66	86.79
gru	reuters	12 months	128	6	0.89	127.26	97.05
gru	dowjones	1 months	64	6	0.89	125.94	89.13
gru	dowjones	3 months	128	4	0.88	131.83	92.54
gru	dowjones	6 months	256	4	0.89	125.57	89.29
gru	dowjones	12 months	128	4	0.89	126.84	91.94
gru	chinaservice	1 months	128	6	0.89	122.44	90.03
gru	chinaservice	3 months	128	4	0.87	138.04	98.17
gru	chinaservice	6 months	128	4	0.88	129.52	91.20
gru	chinaservice	12 months	256	4	0.89	127.25	92.13
gru	no-sentiment	1 months	256	6	0.84	150.85	109.49
gru	no-sentiment	3 months	256	4	0.83	157.38	111.45
gru	no-sentiment	6 months	256	4	0.85	147.14	112.09
gru	no-sentiment	12 months	128	4	0.82	161.15	125.29
lstm	reuters	1 months	128	4	0.89	125.12	87.98
lstm	reuters	3 months	128	2	0.87	135.41	94.14
lstm	reuters	6 months	256	3	0.88	129.96	96.39
lstm	reuters	12 months	256	2	0.88	131.37	92.85
lstm	dowjones	1 months	128	3	0.90	122.05	86.22
lstm	dowjones	3 months	256	3	0.87	135.47	98.11
lstm	dowjones	6 months	64	2	0.88	129.94	95.99
lstm	dowjones	12 months	256	2	0.87	135.88	97.83
lstm	chinaservice	1 months	128	5	0.90	121.83	87.18
lstm	chinaservice	3 months	64	3	0.87	138.20	100.94
lstm	chinaservice	6 months	256	3	0.87	134.75	99.10
lstm	chinaservice	12 months	64	2	0.88	134.19	98.15
lstm	no-sentiment	1 months	128	4	0.90	120.17	84.28
lstm	no-sentiment	3 months	64	2	0.87	134.49	96.65
lstm	no-sentiment	6 months	64	2	0.88	130.91	95.83
lstm	no-sentiment	12 months	256	2	0.87	135.68	97.44
tft	reuters	1 months	128	5	0.88	131.63	95.42
tft	reuters	3 months	128	3	0.88	128.74	90.06
tft	reuters	6 months	128	4	0.88	130.66	92.73
tft	reuters	12 months	256	3	0.88	131.93	93.59
tft	dowjones	1 months	256	4	0.87	134.40	95.62
tft	dowjones	3 months	128	5	0.88	132.15	96.22
tft	dowjones	6 months	256	3	0.87	134.51	98.03
tft	dowjones	12 months	128	4	0.88	134.85	96.60
tft	chinaservice	1 months	64	3	0.88	132.26	98.11
tft	chinaservice	3 months	256	5	0.87	133.48	96.20
tft	chinaservice	6 months	128	2	0.88	130.66	96.20
tft	chinaservice	12 months	64	2	0.88	132.34	96.14
tft	no-sentiment	1 months	128	2	0.88	131.48	94.61
tft	no-sentiment	3 months	64	2	0.88	132.73	93.59
tft	no-sentiment	6 months	64	2	0.88	130.88	96.12
tft	no-sentiment	12 months	256	4	0.88	133.36	96.45

Table 7: Best results (R^2, RMSE and MAE) for the best hyperparameter combination for each model, sentiment source and time window.

Appendix F: Portfolio results for every model, sentiment source and time window combination best grid search combination

Model	Source	Horizon	R^2	RMSE	MAE	HitRate	p_value	TotRet
lstm	chinaservice	1_months	0.90	121.83	87.18	0.56	0.0665	1.16
lstm	chinaservice	3_months	0.87	138.20	100.94	0.56	0.0651	-0.60
lstm	chinaservice	6_months	0.87	134.75	99.10	0.55	0.1760	-0.08
lstm	chinaservice	12_months	0.88	134.19	98.15	0.55	0.1283	-0.21
lstm	reuters	1_months	0.89	125.12	87.98	0.55	0.1594	2.92
lstm	reuters	3_months	0.87	135.41	94.14	0.60	0.0041	0.45
lstm	reuters	6_months	0.88	129.96	96.39	0.53	0.3557	1.09
lstm	reuters	12_months	0.88	131.37	92.85	0.55	0.1283	0.17
lstm	dowjones	1_months	0.90	122.05	86.22	0.56	0.0906	1.36
lstm	dowjones	3_months	0.87	135.47	98.11	0.53	0.3977	-0.39
lstm	dowjones	6_months	0.88	129.94	95.99	0.53	0.3557	0.05
lstm	dowjones	12_months	0.87	135.88	97.83	0.55	0.1693	-0.62
lstm	no-sentiment	1_months	0.90	120.17	84.28	0.58	0.0158	1.31
lstm	no-sentiment	3_months	0.87	134.49	96.65	0.56	0.0890	-0.18
lstm	no-sentiment	6_months	0.88	130.91	95.83	0.58	0.0177	-0.64
lstm	no-sentiment	12_months	0.87	135.68	97.44	0.53	0.3483	-0.42
gru	chinaservice	1_months	0.89	122.44	90.03	0.56	0.0906	1.52
gru	chinaservice	3_months	0.87	138.04	98.17	0.54	0.2035	-0.48
gru	chinaservice	6_months	0.88	129.52	91.20	0.57	0.0534	0.36
gru	chinaservice	12_months	0.89	127.25	92.13	0.54	0.2193	0.05
gru	reuters	1_months	0.89	126.34	89.13	0.54	0.2611	-0.19
gru	reuters	3_months	0.87	135.72	98.21	0.54	0.2035	-0.57
gru	reuters	6_months	0.90	117.66	86.79	0.61	0.0018	2.66
gru	reuters	12_months	0.89	127.26	97.05	0.51	0.7188	-0.71

Continued on next page

Model	Source	Horizon	R^2	RMSE	MAE	HitRate	p_value	TotRet
gru	dowjones	1_months	0.89	125.94	89.13	0.55	0.1594	-0.28
gru	dowjones	3_months	0.88	131.83	92.54	0.55	0.1573	0.26
gru	dowjones	6_months	0.89	125.57	89.29	0.57	0.0377	1.25
gru	dowjones	12_months	0.89	126.84	91.94	0.58	0.0347	-0.15
gru	no-sentiment	1_months	0.84	150.85	109.49	0.55	0.1213	-0.54
gru	no-sentiment	3_months	0.83	157.38	111.45	0.54	0.2588	-0.44
gru	no-sentiment	6_months	0.85	147.14	112.09	0.50	0.9435	0.08
gru	no-sentiment	12_months	0.82	161.15	125.29	0.55	0.1693	-0.52
bilstm	chinaservice	1_months	0.89	125.11	89.85	0.57	0.0478	0.50
bilstm	chinaservice	3_months	0.88	130.69	95.01	0.57	0.0328	0.07
bilstm	chinaservice	6_months	0.89	126.80	96.00	0.57	0.0534	0.33
bilstm	chinaservice	12_months	0.88	130.00	98.25	0.54	0.2193	0.56
bilstm	reuters	1_months	0.89	126.79	91.59	0.55	0.1213	0.70
bilstm	reuters	3_months	0.88	132.66	92.02	0.60	0.0041	2.32
bilstm	reuters	6_months	0.89	126.92	88.74	0.55	0.1344	-0.46
bilstm	reuters	12_months	0.89	128.64	95.47	0.52	0.6141	-0.67
bilstm	dowjones	1_months	0.88	130.12	93.37	0.52	0.5751	0.19
bilstm	dowjones	3_months	0.88	129.93	92.71	0.54	0.2035	-0.34
bilstm	dowjones	6_months	0.89	126.94	88.22	0.58	0.0261	-0.23
bilstm	dowjones	12_months	0.89	123.87	91.57	0.54	0.2193	-0.21
bilstm	no-sentiment	1_months	0.89	124.68	90.66	0.55	0.1213	-0.01
bilstm	no-sentiment	3_months	0.88	131.74	89.92	0.57	0.0328	0.05
bilstm	no-sentiment	6_months	0.89	125.87	89.84	0.55	0.1760	0.75
bilstm	no-sentiment	12_months	0.89	127.39	91.64	0.53	0.4277	0.02
convlstm	chinaservice	1_months	0.90	117.84	82.64	0.58	0.0158	-0.29
convlstm	chinaservice	3_months	0.89	124.18	87.29	0.57	0.0328	-0.11
convlstm	chinaservice	6_months	0.87	134.19	97.63	0.56	0.1007	0.39
convlstm	chinaservice	12_months	0.89	125.63	93.95	0.54	0.2788	0.35
convlstm	reuters	1_months	0.90	117.84	82.64	0.58	0.0158	-0.29
convlstm	reuters	3_months	0.88	132.64	90.43	0.57	0.0467	0.06
convlstm	reuters	6_months	0.87	134.80	96.54	0.56	0.0740	-0.51
convlstm	reuters	12_months	0.88	132.48	97.26	0.55	0.1283	-0.20
convlstm	dowjones	1_months	0.90	117.84	82.64	0.58	0.0158	-0.29
convlstm	dowjones	3_months	0.88	132.64	90.43	0.57	0.0467	0.06
convlstm	dowjones	6_months	0.87	134.80	96.54	0.56	0.0740	-0.51
convlstm	dowjones	12_months	0.88	134.74	97.52	0.56	0.0954	-0.59
convlstm	no-sentiment	1_months	0.90	117.59	82.30	0.59	0.0104	-0.29
convlstm	no-sentiment	3_months	0.88	129.37	89.36	0.56	0.0651	0.22
convlstm	no-sentiment	6_months	0.88	132.81	96.95	0.55	0.1344	0.24
convlstm	no-sentiment	12_months	0.88	130.05	96.46	0.52	0.6141	1.02
tft	chinaservice	1_months	0.88	132.26	98.11	0.57	0.0337	0.05
tft	chinaservice	3_months	0.87	133.48	96.20	0.58	0.0226	-0.39
tft	chinaservice	6_months	0.88	130.66	96.20	0.55	0.1760	-0.58
tft	chinaservice	12_months	0.88	132.34	96.14	0.57	0.0497	-0.41
tft	reuters	1_months	0.88	131.63	95.42	0.56	0.0906	-0.19
tft	reuters	3_months	0.88	128.74	90.06	0.57	0.0328	-0.59
tft	reuters	6_months	0.88	130.66	92.73	0.54	0.2265	0.00
tft	reuters	12_months	0.88	131.93	93.59	0.58	0.0347	-0.35
tft	dowjones	1_months	0.87	134.40	95.62	0.54	0.2058	-0.23
tft	dowjones	3_months	0.88	132.15	96.22	0.55	0.1573	-0.44
tft	dowjones	6_months	0.87	134.51	98.03	0.53	0.4348	-0.55
tft	dowjones	12_months	0.88	134.85	96.60	0.54	0.2193	-0.73
tft	no-sentiment	1_months	0.88	131.48	94.61	0.54	0.2611	-0.58
tft	no-sentiment	3_months	0.88	132.73	93.59	0.57	0.0467	-0.82
tft	no-sentiment	6_months	0.88	130.88	96.12	0.53	0.3557	-0.53
tft	no-sentiment	12_months	0.88	133.36	96.45	0.55	0.1693	-0.52

Table 8: Trading simulation results for the best hyperparameter combination for each model, sentiment source and time window. The p-value here measures the statistical significance of the model's directional accuracy relative to random chance (50%). The TotRet metric shows the total return of the trading strategy over the backtest period.

Flipper: An Extended Document-Level Financial Dataset for Training and Evaluation with Annotated Discourse Phenomena

Mariam Nakhlé[1,2], **Rachel Atherly**[1], **Gabriela González Sáez**[1], **Marco Dinarelli**[1], **Raheel Qader**[2], **Hervé Blanchon**[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP , LIG, 38000 Grenoble, France
(2) Dragon LLM, 75008, Paris, France
contact: mariam.nakhle@univ-grenoble-alpes.fr

Abstract

We present a new resource for Machine Translation (MT), namely a training and evaluation dataset containing parallel documents issued from authentic data in the financial domain. We cover five language pairs: English-French, English-Spanish, English-German, English-Italian and French-Spanish. The total number of parallel documents is 122k and the number of tokens is 118M (source and target combined). MT has improved greatly in recent years, but certain phenomena still cause errors, particularly when context spans beyond a single sentence. Errors can lead to mistranslated pronouns, incorrect gender or number agreement, and inconsistent terminology, which can be especially problematic in high-stakes domains like finance. We therefore construct the dataset at document level (rather than sentence-level) and also produce fine-grained annotations of context-sensitive phenomena. The annotation was performed using preexisting tools and custom scripts. Thus the process can be replicated on different parallel data from other domains. The annotated phenomena are: formality, gender, terminology consistency, verb form and sentence reordering. This aims to improve document-level evaluation of MT models by enabling evaluation solely on texts containing a particular phenomenon of interest. Our primary contribution is the creation and public release of Flipper, a multilingual document-level parallel dataset in the financial domain, designed to support both training and targeted evaluation of context-sensitive machine translation.

Keywords: Financial NLP, Machine Translation, Parallel data, Annotation

Lang. Pair	Documents	Sentences
ENDE	34,438	1,268,547
ENES	12,059	286,376
ENFR	47,327	1,399,169
ENIT	14,449	330,128
FRES	14,588	325,318
TOTAL	**122,861**	**3,609,538**
	Src. tokens	**Trg. tokens**
ENDE	20,351,372	18,347,846
ENES	4,942,426	5,403,083
ENFR	22,119,921	23,811,471
ENIT	5,038,081	5,179,604
FRES	6,206,158	6,253,966
TOTAL	**58,657,958**	**58,995,970**

Table 1: Statistics of the full dataset. For every language pair, 2000 documents are reserved for the evaluation split.

1. Introduction

While Context-Aware Neural Machine Translation (CA-NMT) has received considerable attention in recent years (Toral et al., 2018; Läubli et al., 2020), with numerous works focusing on architectural improvements and modelling strategies, advances in context-aware evaluation have not progressed at the same pace. Many studies still report quantitative results based on sentence-level metrics such as BLEU or COMET (Papineni et al., 2002; Rei et al., 2020), which are not designed to capture document-level phenomena. In parallel, several works rely on ad-hoc evaluation datasets, such as contrastive test suites (Bawden et al., 2018; Müller et al., 2018), that target specific discourse phenomena but do not reflect realistic document-level translation settings.

This imbalance between modelling and evaluation is particularly evident in specialised domains such as finance. Despite the growing importance of domain-adapted MT, very few document-level datasets have been released for the financial domain, and even fewer explicitly target context-sensitive phenomena.

The financial domain constitutes a highly relevant and challenging use case for context-aware MT. It is characterised by specialised terminology, strict regulatory requirements, and a high demand for fast and reliable translations. Providers of financial products are legally required to supply translated versions of documentation in every country where the product is commercialised, creating a strong need for high-quality, in-domain translation systems.

Financial documents exhibit several properties that make document-level evaluation especially crucial. Terminology must remain consistent throughout a document, and the correct translation of a term often depends on definitions introduced earlier in the text—particularly in legal documents, where key entities are explicitly defined at the beginning and must be referred to consistently thereafter. In

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 78–86

addition, numerical consistency is essential: financial reports frequently contain tables with monetary amounts, and inconsistent formatting (e.g., 5 million USD, 5M USD, or $5M) within or across tables is unacceptable. These characteristics highlight the limitations of sentence-level evaluation and underline the need for document-level resources tailored to this domain.

To address this gap, we propose Flipper[1], a new document-level parallel dataset for the financial domain designed for both training and evaluation of CA-NMT systems. We collect a large number of publicly available parallel documents issued by asset management firms. The documents, originally in PDF format, are primarily legal and marketing materials, including annual reports, Key Information Documents, marketing factsheets, manager commentaries, and ESG or sustainability reports.

To convert these documents into parallel data suitable for MT training, we adopt a pipeline inspired by Nakhlé et al. (2025). As in their work, we extract texts from PDFs and align full documents or sections rather than individual sentences, enabling the preservation of higher-level discourse structure and allowing sentence reordering within sections. However, Flipper substantially extends and improves upon this approach. We introduce a more robust alignment method and perform extensive deduplication of near-duplicate content through custom preprocessing steps, resulting in more diverse data, but still a much larger dataset. Beyond data construction, we perform careful annotation of context-sensitive phenomena that are annotated using an inline XML format. The annotation was performed using preexisting annotation tools and custom-made scripts. The annotated phenomena are: formality, gender, terminology consistency, verb form and sentence reordering.

Unlike the DOLFIN dataset, Flipper includes both training and evaluation splits, which makes it a valuable resource for model training. 2000 documents are reserved for the evaluation set, we deliberately select sections with a strong presence of discourse- and context-dependent phenomena. This design positions our evaluation split between a standard test set and a targeted test suite: it remains authentic and document-based, yet it increases the density of challenging phenomena, making evaluation more informative.

Flipper covers five language pairs: English–Italian (En–It), English–Spanish (En–Es), English–German (En–De), English–French (En–Fr), and French–Spanish (Fr–Es), the latter being the only non-English pair. As stated in Table 1, the total number of parallel sections is 122k and the number of tokens is 118M (source and target tokens combined).

The dataset can be used to train or fine-tune document-level MT systems or Large Language Models (LLMs) and to conduct targeted evaluation of specific context-sensitive phenomena through its fine-grained inline annotations. To the best of our knowledge, Flipper is the first publicly available parallel dataset in the financial domain that simultaneously satisfies three key criteria: (1) document-level structure, (2) domain-specific coverage of authentic financial texts, and (3) suitability for both training and targeted evaluation. Our contributions can be summarized as following:

- The creation and public release of a document-level parallel dataset for the financial domain;
- The provision of a dedicated training split tailored to document-level MT;
- The targeted selection of evaluation data enriched with discourse- and context-sensitive phenomena;
- Fine-grained inline annotation of multiple context-sensitive phenomena, including formality, gender, terminology consistency, verb form, and sentence reordering.

2. Related work

Natural Language Processing (NLP) for finance. The use of NLP techniques across various applications has grown significantly in recent years across a broad range of domains, and finance is no exception. Some of the well-studied tasks in this field include named entity recognition (Salinas Alvarado et al., 2015), question answering (Chen et al., 2021; Maia et al., 2018), sentiment analysis (Malo et al., 2014), and topic modeling (Jehnen et al., 2025).

Several finance-oriented language models have also been developed. These include encoder-only models such as FinBERT (Araci, 2019), as well as decoder-only models such as BloombergGPT (Wu et al., 2023), FinMA (Xie et al., 2023), FinGPT (Wang et al., 2023), and LLM Pro Finance (Caillaut et al., 2025). These models can be applied to a variety of downstream tasks, including Machine Translation, which is the focus of this work.

Machine Translation for finance. Although Machine Translation (MT) is one of the core tasks in NLP, resources tailored to the financial domain remain scarce. With regard to financial-domain resources, the diachronic banking magazine collections (Volk et al., 2016) provide relevant material; however, the data is available only at the sentence level. Nakhlé et al. (2025) present a document-level parallel dataset for the financial domain, but it is limited to evaluation data. More generally, several document-level datasets have been proposed for

[1] https://huggingface.co/datasets/DragonLLM/Flipper

training purposes (Koehn, 2005; Tiedemann, 2012; Cettolo et al., 2012; Lison and Tiedemann, 2016; Wicks et al., 2024). As for evaluation data, the majority of available benchmarks remain sentence-level, such as the datasets released annually within the WMT conference (Kocmi et al., 2025). Some test sets do include document boundaries; however, even in these cases, the primary unit remains the sentence, meaning that individual sentences are aligned within documents.[2] Evaluation typically measures the overall quality of a model by feeding translations to the metric one sentence at a time. By design, this excludes looser translation strategies, such as sentence reordering or splitting. In contrast, our approach aims to accommodate such phenomena, which is why the main unit of our proposed dataset is the document rather than the sentence.

Another line of work in context-aware evaluation involves targeted test suites that focus on linguistic phenomena requiring context beyond a single sentence for correct evaluation. These resources are often constructed manually and are designed to probe systems with respect to specific discourse-level challenges. While they provide valuable insights into particular phenomena, they remain limited in scope, language coverage, and applicability to overall translation quality.

In this work, we aim to address the gap in available resources by providing parallel data that meet three criteria: (1) document-level structure, (2) coverage of the financial domain, and (3) suitability for both training and evaluation. Our dataset, Flipper, enables the training and evaluation of MT models on challenging texts drawn from authentic financial documents containing context-sensitive phenomena, thereby combining general quality assessment with targeted evaluation.

3. Dataset collection

Our dataset collection procedure is made of the following processing steps: 1) PDF-to-text (Markdown) extraction, 2) noisy data filtering, 3) alignment of sections within documents, 4) near-duplicate removal, and 5) context-sensitive phenomena annotation. This pipeline is inspired by the one presented in Nakhlé et al. (2025), as we similarly worked with financial PDF documents and aimed to produce a document-level dataset. However, we introduce improvements in certain aspects, particularly in the alignment approach, which is crucial for parallel data. We also modified the deduplication process, as the data still appeared to be highly repetitive. In contrast to the cited previous work, we do not perform quality estimation filtering, as we believe it introduces bias into the data curation process. Finally, we produced fine-grained inline annotations of context-sensitive phenomena.

Alignment. We used a different alignment approach, as the original method was rather naive (two sections were considered aligned if their first and last sentences were aligned). In our approach, we use the same aligner LASER (Schwenk and Douze, 2017), but we modified the decision logic for determining whether two documents are aligned.

We compute LASER alignment scores between each source sentence and a window of size m in the target document. For a source sentence with index i, the target window ranges from `i - m/2` to `i + m/2`. We then compute an average score for the entire section based on the scores of all source sentences, taking the maximum score for each source sentence.

To determine an acceptability threshold, we manually inspected 1,200 sections. Many collected sections were challenging cases: although they appeared similar, they were not faithful translations of one another. A threshold of 0.80 was selected as the best balance between permissiveness and strictness, while retaining the maximum number of correctly aligned sections.

Deduplication. Another limitation of the dataset described by Nakhlé et al. (2025) is its repetitiveness, which is directly related to the financial domain. Companies are legally required to publish certain documents periodically; for efficiency, these documents often contain repeated sections. Additionally, some documents—such as Key Information Documents—must present information in a standardized format. As a result, texts describing different funds are often identical, differing only in numerical values.

To address this issue, we applied stricter deduplication using the MinHash algorithm implemented in the *text-dedup* library by Mou et al. (2023), complemented by a preprocessing step. Specifically, we tokenized the texts and masked numerical values so that texts differing only in numbers would still be identified as duplicates and removed. We then set the deduplication threshold to 0.5. For this step, we considered merged source and target documents.

4. Annotation

In order to enrich the dataset with more nuanced information, we performed an annotation of context-sensitive phenomena. These are linguistic phenomena that require extra-sentential information to be correctly interpreted and translated, and they present a particular challenge for Machine Translation, especially when models operate on isolated sentences. This annotation enables future users

[2] To the best of our knowledge, the WMT test sets released in 2025 are the first ones in which the document is treated as the main unit.

of the dataset to target specific phenomena by extracting the relevant sections and analysing model performance accordingly.

To carry out the annotation, we employed several pre-existing tools designed for this purpose. Formality, plurality, and grammatical gender are among the more common context-sensitive phenomena and were annotated using existing tools developed for parallel data annotation. To further diversify the range of annotated phenomena, we developed custom tools to address two additional aspects: terminology consistency, which ensures that terms are translated consistently throughout a document, and high-level information reordering, which captures changes in sentence order during translation.

4.1. Annotation format

The annotations are provided in an inline XML format. This standardises the notation across different tools and annotation methods, enabling multiple phenomena to be marked within the same section.

The main tag used is `<annotation>`, which includes a `tool` attribute indicating the source of the annotation. The `phenomenon` attribute specifies the type of context-sensitive phenomenon being marked. Depending on the phenomenon, additional attributes provide more detailed information. Below are two examples of annotations, illustrating the formality phenomenon and the terminology consistency phenomenon, respectively.

> Before buying or switching Units, <annotation tool="ctxpro" rule="NOM.FORM+PLUR" phenomenon="formality"> you </annotation> should read the relevant KIID.

> The absolute and/or relative returns shown in the performance attribution section of this document (hereafter <annotation tool="custom" phenomenon="terminology consistency" id="def_1" refersTo="performance attribution section"> Reporting </annotation>) may differ from the returns in other statements or reports provided due to the use of different methodologies. For official use, only the official statements and reports should be used and not the figures provided in this <annotation tool="custom" phenomenon="terminology consistency" id="ref_1" refersTo= "def_1"> Reporting </annotation>.

4.2. Annotations using preexisting tools

4.2.1. CTXPRO

CTXPRO (Wicks and Post, 2023) is an automatic annotation tool that identifies phenomena that require contextual information to be translated correctly. The identification process is rule-based and relies on linguistic information specific to the language pair. The pipeline takes sentence-aligned parallel data as input and outputs information indicating which segments contain ambiguities.

CTXPRO was run on all language pairs except French–Spanish, as the tool requires at least one of the languages to be English. It was able to tag phenomena related to gender and formality ambiguities. The formality phenomenon arises when one language uses an ambiguous pronoun whose translation into the target language depends on the level of formality. For example, the English pronoun “you“ translates into French as “tu“ (informal) or “vous“ (formal). The gender phenomenon occurs with pronominal anaphora, where correct translation of a pronoun requires access to its antecedent. For instance, the English pronoun “it“ may translate into French as “il“ (masculine) or “elle“ (feminine), depending on the gender of the antecedent in the translation.

In the annotation, we include the rule as provided by the tool via the `rule` attribute. An example of a rule is NOM.FORM.SING. We use the middle element of the notation to indicate the phenomenon category, in this case formality.[3]

4.2.2. MuDA

The second pre-existing annotation tool employed was the Multilingual Discourse-Aware (MuDA) benchmark (Fernandes et al., 2023). Similarly to CTXPRO, MuDA identifies and annotates phenomena that require context spanning multiple sentences or longer stretches of text. The tool shares the same limitation of requiring English to be part of the language pair. On our data, the phenomena identified by this tool are verb form and formality. The verb form phenomenon arises when the choice of a verbal form (such as tense or mood) depends on extra-sentential context, as is the case in text passages containing a succession of verbs in the imperfect tense.

4.3. Hand crafted annotations

4.3.1. Terminology consistency

Financial documents contain numerous terms that are defined or redefined at a specific point in the

[3]We refer the reader to the original paper for a detailed description of the rules.

text. For example, in the sentence “the total risk is measured and checked using the relative value at risk (hereinafter ‘relative VaR’) method,“ the term “relative VaR“ must remain unchanged throughout the rest of the document during translation. A similar situation arises with abbreviated company names. In our financial documents, it is common for the full name or legal designation of a financial fund to be stated at the beginning and subsequently referred to as “the Fund” throughout the remainder of the text. This creates terms that require consistent and accurate translation, even though their definition appears only once, several sentences or even paragraphs earlier. This phenomenon is particularly critical in financial and legal contexts, where the acceptable margin of error is significantly lower than in other domains.

To identify instances of terminology consistency, a semi-manual program was designed and implemented. First, regular expressions were used to search the dataset for the trigger words “hereafter” and “hereinafter,” which commonly signal the introduction of a definition, as illustrated in the example above. The term introduced as the new definition (referred to as the “alias”) was then identified using an additional set of regular expression patterns. However, due to numerous edge cases and variations in wording, identifying the term being redefined (the “head”) required more manual effort. While the pattern could detect potential instances of terminology consistency, accurately annotating the details required further validation. Therefore, an interactive interface was developed that used the regular expressions to display the predicted “alias” and “head” pairs individually for review. Using this approach, the phenomenon was annotated with human oversight to approve, reject, or edit each tag. This method enabled efficient annotation while increasing confidence that the final dataset did not contain erroneous tags resulting from unique edge cases.

This phenomenon includes attributes `id` and `refersTo` in the annotation tags. The `refersTo` attribute links a term to its corresponding definition. The `id` attribute can take the form `def_x` or `ref_x`, depending on whether the tag marks the definition of a terminology rule or a subsequent reference to it. The attribute `id="def_x"` appears within hereafter declarations, while `refersTo` contains the word or phrase introduced by the definition. If the tag marks a later reference to an existing definition, `refersTo` contains only the corresponding definition ID. This structure makes it possible to count how many references a given terminology rule has. Definition IDs are global (e.g., there is only one `def_1`), whereas reference IDs are local; for instance, `ref_1` denotes the first reference associated with each `def_x`.

4.3.2. Sentence reordering

During translation, a certain degree of high-level information reordering may occur. It is not always possible to translate sentences word for word or to process documents strictly sentence by sentence. In some cases, sentences must be reordered, split, or merged to produce a more natural and fluent text in the target language. This poses a particular challenge for MT evaluation, as the source and target texts may not preserve sentence-to-sentence alignment, while evaluation metrics typically process sentences individually. For this reason, we sought to annotate this phenomenon in order to enable targeted evaluation and to analyze the extent to which it occurs in the translation of financial documents.

To annotate sentence reordering, we first preprocessed the texts by removing tables and other heavy *Markdown* formatting. The sentences were then encoded using a Sentence-BERT model (Reimers and Gurevych, 2019)[4]. We computed a cosine similarity matrix between source and target sentences to identify the best match for each sentence. If the best match for a given source sentence had a different index in the target text than in the source, it was flagged as a potential reordering. This method detects both direct swaps and broader shifts in sentence order. In several sections, for example, the first sentence is split in translation, shifting all subsequent sentence indices by one throughout the section.

The annotation tag for this phenomenon includes the attributes `from` and `to`, which indicate the original and matched sentence indices, respectively. The `from` attribute refers to the index of the source sentence and, by extension, to the expected index of its corresponding translation. The `to` attribute indicates the index of the target sentence identified as the best match. For example, `from=2 to=3` means that source sentence 2 has a higher cosine similarity with target sentence 3 than with target sentence 2, suggesting that a reordering has occurred.

However, manual analysis of the annotations revealed that this method is insufficient for reliably detecting sentence reordering. In many cases, shifts in sentence indices were caused by errors introduced by the sentence splitter or by minor misalignments between source and target texts. For instance, if the first sentence is omitted in translation, all subsequent sentence indices are shifted. Although we retain the existing annotations, further investigation of this phenomenon is left to future work.

[4]HuggingFace identifier: sentence-transformers/distiluse-base-multilingual-cased-v1

4.4. Annotation statistics

Table 2 presents the results of the annotation over the full dataset. The highest number of annotations is observed for the English–French language pair, indicating that the tools are most effective for this high-resource pair. A notable limitation of the annotation process concerns the only non-English-centric language pair, French–Spanish, for which none of the pre-existing tools were suitable. The most frequently annotated phenomenon is verb form, as identified by the MuDA tool, followed by sentence reordering and formality. Manual inspection showed that some of the annotations point to phenomena that are in fact translatable even without context (despite the tools being specifically designed for this goal) and we suspect that there are context-sensitive phenomena that were missed by the tools and exist in the dataset without an annotation. This imbalance highlights the inherent difficulty of constructing targeted test sets from authentic documents that present specific context-sensitive translation challenges.

5. The training and evaluation splits

Finally, the dataset was divided into training and test splits. The test set contains ten thousand sections, with two thousand sections per language pair. These sections were selected based on the quantity and variety of annotation tags in order to maximize phenomenon diversity and thereby create a more challenging evaluation set, with the aim of reserving the most complex sections for the test set. The training set consists of all remaining sections, with a target size of at least ten thousand aligned sections per language pair.

6. Experiments

6.1. Setup

Next, we conducted an experiment by fine-tuning the 1B-parameter Gemma 3 Instruct model (hereafter gemma-3-1b-it) (Team, 2025). We selected this model due to its broad language coverage, and because our preliminary experiments indicated strong performance on machine translation.

We fine-tuned the model on the newly constructed Flipper dataset using full-parameter supervised fine-tuning implemented with the TRL (Transformer Reinforcement Learning) library. We use the chat template applicable for this model and we add the following prompt via the user content: `Translate the following paragraph from {source_language} to {target_language}.`
`n Do not add anything else at all.{source_text}.` The assistant reply is the translation alone.

The loss was computed over the completion tokens only; in our setup, these correspond to the translated text. Training is conducted for 1 epoch with a batch size of 8, learning rate 1e-4, linear scheduler, and paged AdamW 8-bit optimizer. We use bfloat16 precision with gradient checkpointing enabled and a maximum sequence length of 1024 tokens. Training is performed on a single GPU with seed 42.

6.2. Evaluation

For evaluation, we translated the evaluation split of Flipper and compared the fine-tuned model against the base gemma-3-1b-it model in a zero-shot setting. We report results using the `wmt22-comet-da` metric (Rei et al., 2022).

Table 3 presents the results per language pair, while Table 4 details the results per phenomenon. As shown, fine-tuning improves translation quality for three out of the five language pairs. The largest gain is observed for English–Spanish, followed by English–French and French–Spanish. In contrast, English–German and, to a lesser extent, English–Italian show a slight degradation compared to the base gemma-3-1b-it model.

At the phenomenon level, four phenomena out of five show gains (Formality, Terminology Consistency, Verb Form, and Reordering), with the largest gain observed for Formality and in Verb Form. These gains suggest that the additional training particularly benefits controlled linguistic phenomena. In contrast, performance on Gender decreases, indicating that improvements are not uniformly distributed across all the phenomena.

LLMs typically benefit from exposure to diverse tasks and domains, which help boost performance on the target task (Alves et al., 2024). This may explain the slight degradation observed. Fine-tuning on a narrower dataset or a single task can lead to performance gains in some directions while slightly degrading others, which is why usually the post training phase includes a mix of tasks. A more comprehensive fine-tuning strategy involving multiple tasks and datasets could yield more uniform improvements across languages and phenomena. However, conducting a large-scale, multi-task optimization was beyond the scope of this work. The primary objective of this paper is to introduce the new dataset and demonstrate its usability through a focused fine-tuning experiment, rather than to exhaustively optimize model performance.

Phenomenon	En-Fr	En-Es	En-It	Fr-Es	En-De	Totals
Verb form	88 723	2 332	551	0	0	91 606
Gender	2 203	30	22	0	609	2 864
Formality	6 628	360	481	0	7 513	14 982
Term. consistency	1 360	284	409	0	1 340	3 393
Sent. reordering	6 833	2 512	2 726	2 358	6 611	21 040
Totals	105 747	5 518	4 189	2 358	16 073	133 885

Table 2: Results of the annotation. We report the number of annotation tags for all phenomena (meaning multiple tags can be present in one document), except for the sentence reordering where we report the number of annotated documents, since one shift can cause all the subsequent sentences to be annotated as reordered.

Lang	gemma-3-1b-it	Fine-tuned
En-De	69,77	67,81
En-Es	77,90	80,25
En-Fr	75,06	75,91
En-It	74,30	73,73
Fr-Es	78,19	78,87

Table 3: Comet scores per language pair.

Phenomenon	gemma-3-1b-it	Fine-tuned
Formality	59,22	61,16
Gender	77,33	75,62
Term. consist.	74,69	75,23
Verb form	74,07	75,86
Reordering	67,00	67,64

Table 4: Comet scores per phenomenon.

7. Conclusion

We presented Flipper, a multilingual, document-level parallel dataset for the financial domain, designed to support both training and evaluation of context-aware machine translation. The dataset covers five language pairs and contains 122k parallel sections with 118M tokens. Our work makes several key contributions: We improved data processing pipeline by improving the section-level alignment, and deduplication process, enhancing the quality and diversity of parallel data compared to prior work. We also added inline fine-grained annotation of context-sensitive phenomena, including formality, gender, terminology consistency, verb form, and sentence reordering, enabling targeted evaluation of challenging translation issues. The dataset has a training and evaluation utility since it includes a dedicated training split and a carefully selected evaluation set enriched with discourse- and context-dependent phenomena, making it suitable for both model training and more informative evaluation. By combining authentic financial documents, document-level structure, and rich annotations, Flipper offers a valuable resource for advancing document-level MT in high-stakes, domain-specific settings.

8. Limitations

Text extraction from PDFs remains a bottleneck and can introduce errors that propagate throughout the processing and annotation pipeline. The French–Spanish pair was particularly affected by tool dependencies on English, highlighting the need for further work on annotations, as existing tools proved largely incompatible. Additionally, our approach may misinterpret sentence reordering due to errors from sentence splitting or minor source–target misalignments, which were not fully addressed in this study. These limitations may be subject of future work.

9. References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Gaëtan Caillaut, Raheel Qader, Jingshu Liu, Mariam Nakhlé, Arezki Sadoune, Massinissa Ahmim, and Jean-Gabriel Barthelemy. 2025. The

llm pro finance suite: Multilingual large language models for financial applications.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Simon Jehnen, Joaquín Ordieres-Meré, and Javier Villalba-Díez. 2025. Fintextsim: Enhancing financial text analysis with bertopic.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv:1810.02268*.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. DOLFIN - document-level financial test-set for machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5544–5556, Albuquerque, New Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Martin Volk, Chantal Amrhein, Noëmi Aepli, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9876–9890, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance.

TranslateGemma for ES-EN Financial Reports: Exploring Adaptability to Variable-Sized Contexts

Yanco Amor Torterolo Orta, Melina Chatzi, Antonio Moreno-Sandoval
UAM, UNED
{yanco.torterolo, melina.chatzi}@estudiante.uam.es, antonio.msandoval@uam.es

Abstract

This paper explores bidirectional financial Machine Translation (MT) between Spanish and English, focusing on the specialized domain of annual reports from IBEX 35 companies. Fine-tuned models are compared against zero-shot scenarios through a series of experiments, testing factors such as prompting strategies and model size. On the one hand, this work studies a combination of existing fine-tuning strategies aimed at improving the adaptability of MT models to variable-sized contexts, and, on the other hand, it analyzes the limitations detected in current evaluation metrics. Results are mixed: fine-tuned models show an improvement in both short and long-context scenarios in traditional metrics, while zero-shot predictions are clearly favored by neural metrics. In fact, reference-free assessment of the source and the human reference received worse scores than the off-the-shelf prediction models. Consequently, fine-tuning on the human-made dataset hardly improves the neural metrics against zero-shot generations. This suggests that neural metrics tend to favor the fluency of MT generations and literalness over creativity, among other technical limitations regarding long-context adaptability. From a practical standpoint, the low Translation Edit Rate (TER) scores suggest that specialized fine-tuning remains the most viable path for companies to implement efficient Machine Translation Post-Editing (MTPE) workflows, given the stylistic alignment.

Keywords: financial machine translation, translategemma, long-context evaluation, annual reports

1. Introduction

Companies all over the world use translation as a means of reaching a broader audience beyond their borders. As highlighted by Herrero Rodes and Román Mínguez (2015), these companies draft their annual reports and have them translated into other languages. This process is vital for their business strategy since it allows them to address potential shareholders. Spain is no different, with IBEX 35 companies publishing their annual reports on their websites (Quesada and Espada, 2024).

Natural Language Processing (NLP) tasks have evolved dramatically in the past few years, and MT is no exception. Generative LLMs offer increasingly larger context windows, which is beneficial for nearly every kind of NLP task. This is especially true of financial MT, even reaching document-level (Wang et al., 2023), as the context usually contains elements key to the target text.

This paper explores how the context window affects Machine Translation (MT) and how different metrics respond to different context sizes. Inspired by several existing works and methodologies (Tiedemann and Scherrer, 2017; Johnson et al., 2017; Ding et al., 2021), a Data Augmentation (DA) strategy consisting of duplicating the dataset with a version made of concatenated Translation Units (TU), combined with a bidirectional, variable-sized exposure to the dataset during fine-tuning is tested. It is aimed at improving adaptability to long contexts while assessing the viability of current metrics. To this end, a compact, local model was used for bidirectional financial ES-EN MT. This choice stems from data privacy and environmental footprint concerns, which are key considerations for companies. The democratization of LLMs on consumer-grade hardware is also promoted.

This research focuses on the annual reports of IBEX 35 companies. A parallel corpus consisting of aligned ES-EN TUs was used. It was compiled from 34 pairs of annual reports, amounting to 41,951 TUs of varying sizes—ranging from titles and lists to paragraphs. The word count totals 1,257,458 (ES) and 1,098,426 (EN). A selection of this dataset is available in the following repository: Moreno-Sandoval et al. (2025).

Several research questions will be addressed: (1) Is it worth fine-tuning models on the annual reports of IBEX 35 companies, or similar contexts, given the zero-shot capabilities of current LLMs for MT? (2) From a corporate perspective, does local MT provide an acceptable draft as a starting point for human translators? (3) Does this variable-sized context strategy effectively improve scores compared to fine-tuning on a default dataset? (4) Are there limitations in current metrics?

The rest of the paper is structured as follows: Section 2 reviews related work; Section 3 lays the groundwork for the experiments; and Section 4 provides a detailed analysis of the results. Finally, in Section 5, conclusions are drawn based on the findings.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 87–97

2. Related work

The most similar work found examines Arabic (AR-EN) financial MT (Alghamdi et al., 2023), where it is noted that off-the-shelf neural MT models exhibit an inability to translate domain-specific texts. However, LLMs have improved significantly since 2023.

In this regard, generative LLMs are being increasingly used, as recent iterations of the WMT suggest (Kocmi et al., 2023, 2024, 2025). This trend applies to both translation and evaluation, which is highly relevant to the present work, since several LLMs are used for evaluation and TranslateGemma—a recently released state-of-the-art (SOTA) model—is employed. The subtitles of these findings reports illustrate the rapid evolution of LLMs: *2023: LLMs Are Here But Not Quite There Yet*, *2024: The LLM Era is Here but MT is Not Solved Yet*, and *2025: Time to Stop Evaluating on Easy Test Sets*.

Regarding long-context settings, several studies address document-level MT, such as (Herold and Ney, 2023; Wang et al., 2023). Furthermore, this work provides a document-level financial test set for MT (Nakhlé et al., 2025). A notable study analyzing MT metrics can be found in (Di Natale et al., 2025).

In general, NLP tasks targeting financial texts have gained significant attention, as shown in works such as (Ke et al., 2025). Additionally, there are benchmarks like WMT24++ (Deutsch et al., 2025) that comprise several domains—literary, news, social, and speech—but lack a specific finance component. A more adequate benchmark would be TransBench (Li et al., 2025), as it includes specialized subdomains such as e-commerce. In this paper, no benchmarks were used due to the task-specific nature of the dataset; however, they remain a valuable consideration for future research.

3. Settings and experiments

3.1. Hardware settings and environment

All experiments described in this paper were conducted on a system equipped with an `AMD Ryzen 7 9800X3D` processor, an `NVIDIA RTX 5080` GPU with 16GB of VRAM, and 32GB of DDR5 RAM. For fine-tuning and inference, the official `Unsloth` Docker image was used to prevent dependency issues. Metrics were computed within a standard Conda environment.

3.2. Models and dataset

The models utilized in this study consist of two variants of the TranslateGemma family: `google/translategemma-4b-it` and `google/translategemma-12b-it` (Finkelstein et al., 2026). Based on the Gemma 3 architecture, these models were specifically fine-tuned by Google to excel in MT across 55 languages. As instruction-tuned variants, they are designed to follow prompts effectively.

Regarding the dataset, the TUs from the corpus were already segmented into paragraph-level structures, although some fragments were short (e.g., titles and lists). Two different configurations were evaluated: (a) the **original version**, which uses TUs not exceeding 697 tokens (including both source and target language references); and (b) the **mixed version**, which combines the original dataset with a reformulated version of itself, where the same TUs are concatenated into larger segments of up to 2,102 tokens. This automated chunking process sought to obtain larger TUs while adhering to rules designed to preserve section and list integrity within each chunk. The original TUs remained intact in the process, as they were not split. Table 1 summarizes the characteristics of each dataset. Each entry includes an ID, the source language (Spanish), and the target language reference (English).

dataset	train	val	test	total
original	39,857	1,047	1,047	41,951
mix	43,738	1,169	1,169	46,076

Table 1: Number of TUs in each dataset split.

It is worth noting that the context window for TranslateGemma is 2,000 tokens. While the model card lists input and output separately, this typically encompasses the source text, prompt, and target output. However, this is a functional limit established during Google's translation-focused fine-tuning. Including the prompt, the mixed dataset contains TUs reaching 2,300 tokens. Since the underlying Gemma 3 architecture supports a context window of up to 128,000 tokens, and the objective of this work is to further fine-tune the models, this should not hinder performance.

3.3. Variable-sized context adaptability

By randomly interleaving individual TUs with concatenated sequences from the mixed dataset, the model was exposed to varying contextual scales. This approach mitigates the "sentence-level bias" (or short-sequence bias) typical of standard MT datasets and encourages discursive consistency across extended financial narratives. Since preceding and subsequent contexts are often crucial for the accurate translation of a segment, leveraging an expanded context window is potentially beneficial. Furthermore, this method serves as a data augmentation technique by offering alternative ways of presenting the training data. This randomized interleaving prevents length-related bias and

promotes flexibility regarding context size. Complementing this, bidirectionality is a core aspect of this strategy. The model is exposed to the dataset for both translation directions (ES-EN and EN-ES). The random sampling of translation directions during training prevents directional bias and results in a more robust system.

3.4. Chat template and prompt

The default inference process in TranslateGemma relies on a complex, hardcoded Jinja2 chat template embedded within the model's tokenizer, which acts as a rigid preprocessing layer. This template, hereafter referred to as **GP**, contains an extensive internal mapping of ISO language codes and enforces a strict conversational structure that includes system role validation and specific token formatting, such as start-of-turn and end-of-turn indicators. While this ensures adherence to Google's official specifications, it introduces a significant computational bottleneck and unnecessary token overhead. In contrast, the official Ollama adaptation[1], referred to as **GPO**, utilizes a more straightforward prompting strategy that bypasses the heavy Jinja2 logic in favor of a direct instructional format. GPO strips away the structural constraints of the hidden template and delivers the core translation task directly to the model. The GPO approach is provided in Figure 1. Effectively, the underlying information and the translation intent remain the same—what varies is the efficiency of the delivery and the reduction in preprocessing latency.

The GP and GPO chat templates are compared alongside a third prompt variant, which consists of a minimal instruction:
`You are a professional translator. Translate the following text from Spanish to English` (or *vice versa*).

> You are a professional {SOURCE_LANG} ({SOURCE_CODE}) to {TARGET_LANG} ({TARGET_CODE}) translator. Your goal is to accurately convey the meaning and nuances of the original {SOURCE_LANG} text while adhering to {TARGET_LANG} grammar, vocabulary, and cultural sensitivities.
>
> Produce only the {TARGET_LANG} translation, without any additional explanations or commentary. Please translate the following {SOURCE_LANG} text into {TARGET_LANG}:
>
> {TEXT}

Figure 1: GPO prompting implementation by Ollama based on Google's implementation (GP).

[1] https://ollama.com/library/translategemma

3.5. Fine-tuning and inference

Several fine-tuning and inference components were individually analyzed to conduct a comprehensive ablation study. Given the exceptional multilingual capabilities of current LLMs, the inclusion of non-fine-tuned (zero-shot) baselines was deemed essential. Consequently, these models were evaluated in an inference-only setup to establish a comparative benchmark. Furthermore, multiple inference variables were examined, such as the application of **beam search** and the impact of model size. Each translation direction was also analyzed, enabling cross-directional comparisons. The 16GB VRAM constraint necessitated a resource-aware experimental design incorporating QLoRA and careful hyperparameter selection, in which the use of the `Unsloth` library proved instrumental.

3.6. Hyperparameters

Fine-tuning hyperparameters using `Unsloth` remained identical across experiments, with minor adjustments between the 4B and 12B model versions. Both models were loaded in 4-bit precision with a per-device training batch size of 2 and 8 gradient accumulation steps. The evaluation batch size was set to 1, while `eval_accumulation_steps` were set to 4 for the 4B version and 1 for the 12B version to prevent out-of-memory (OOM) errors during evaluation. Training was governed by an early-stopping mechanism (monitoring *eval loss*) with a patience of 3 evaluation calls, occurring every 500 steps. Although 5 epochs were initially scheduled, the models reached early stopping between 2.2 and 4.1 epochs, depending on the dataset, prompt, and model size. The best model (lowest validation loss) was consistently loaded upon completion. Additional parameters included a learning rate of 1×10^{-5}, 400 warmup steps, and the `paged_adamw_8bit` optimizer with a linear scheduler. Gradient checkpointing was enabled (`use_reentrant=False`). Regarding QLoRA configurations, a rank (r) of 16 and a LoRA alpha (α) of 16 were applied to all linear layers.

3.7. Metrics

The metrics chosen for this work are summarized in Table 2. MetricX and XCOMET are established as SOTA metrics, as evidenced by their continued adoption in the most recent WMT25 shared task (Juraska et al., 2025). However, MetricX-24 (Juraska et al., 2024) was utilized instead of the 2025 version, as the latter had not been fully released at the time of writing. According to Juraska (Juraska, 2025), the 2025 iteration "did not outperform other fine-tuned metrics on the WMT25 test set" and "didn't provide a consistent improvement

over MetricX-24." Consequently, they recommend adhering to the MetricX-24 version.

metric	model
MetricX	google/metricx-24-hybrid-large-v2p6-bfloat16
MetricX_QE	google/metricx-24-hybrid-large-v2p6-bfloat16
MetricX_REFQA	google/metricx-24-hybrid-large-v2p6-bfloat16
XCOMET	Unbabel/XCOMET-XL
XCOMET_QE	Unbabel/XCOMET-XL
XCOMET_REFQA	Unbabel/XCOMET-XL
CHREF++	n/a
TER	n/a
BLEU	n/a

Table 2: Metrics and models employed.

MetricX-24 utilizes a hybrid transformer-based architecture—leveraging large-scale encoder-decoder models like mT5—to assess translation quality, consistently demonstrating superior correlation with human judgment (Freitag et al., 2024). The metric operates by encoding the source, reference, and prediction into a shared embedding space, where it performs reference-based regression trained on Multidimensional Quality Metrics (MQM) data. This allows the model to move beyond surface-level lexical overlaps and evaluate semantic fidelity. As a regressive metric trained to predict human-assigned error scores, it operates on an inverse scale where lower values denote higher translation quality.

Complementing this, **XCOMET** provides an error-aware evaluation by integrating a multi-task learning objective into the COMET framework. By utilizing cross-lingual encoders like XLM-RoBERTa and training on both Direct Assessment (DA) and MQM data, it distinguishes stylistic nuances from critical semantic errors to produce a normalized quality score. To overcome the encoder's 512-token constraint in long-form financial documents, a hierarchical dynamic chunking strategy was implemented. This methodology prioritizes line-by-line alignment, followed by sentence-level segmentation and a length-based fallback to ensure comprehensive coverage. Although averaging segment scores provides a global quality estimate, this fragmentation introduces specific risks: the potential loss of cross-chunk cohesion, the risk of alignment drift during fallback splitting, and the dilution of localized critical errors within the aggregate mean.

As a fundamental component of the evaluation framework, an analysis of the intrinsic quality of the gold standard references was performed to determine whether the human-provided translations effectively represent a definitive upper bound for model predictions. This comparative analysis was facilitated by the dual-mode architecture of MetricX and XCOMET, which support both reference-based (source, reference, and prediction) and reference-free evaluation. Specifically, two dimensions were assessed: (a) an evaluation of the source and the reference, hereafter referred to as Reference Quality Assurance (**RefQA**), and (b) an evaluation of the source and the model's prediction, referred to as Quality Estimation (**QE**). Comparing these two dimensions enables the identification of whether a specific score reflects suboptimal model performance or stems from underlying inconsistencies within the reference material itself (Freitag et al., 2023).

Besides neural metrics, **chrF++** and **TER** (Translation Edit Rate) were included to evaluate different aspects of the translation. While chrF++ measures character-level accuracy, TER estimates the effort a human would need to correct the text. **BLEU** metric was also implemented via the SacreBLEU toolkit as a standard benchmark. Even though BLEU has limitations in capturing full meaning, it remains one of the most widely used metrics in the field, enabling comparability with broader research. Regarding their interpretation, higher chrF++ and BLEU scores, and lower TER values denote superior translation quality. These traditional metrics are more transparent and provide a reliable baseline to check if the main neural metrics suffer from metric bias. This phenomenon is an instance of reward hacking , where the utility function (the reward model) improves but the system's behavior diverges from actual quality or human preferences (Kovacs et al., 2024).

4. Results and discussion

4.1. GP vs GPO

The impact of the chat template on model behavior was initially evaluated, as it informs several subsequent experimental decisions. Table 3 provides a performance and efficiency comparison between the GP and GPO approaches. Both configurations were tested using `google/translategemma-4b-it` in a zero-shot setup. To prioritize inference speed, all 4B models were loaded in 16-bit precision. While the difference in translation quality remains negligible, the efficiency gains are substantial: GPO achieves a reduction in both execution time and energy consumption of over 98% relative to the standard GP template.

It is important to clarify that the GPO version was executed via Ollama, whereas the GP version utilized the `Unsloth` fast inference framework. Given that `Unsloth` is highly optimized for speed, the disproportionate latency observed in

Metric	4b_gp	4b_gpo
Translation Metrics		
MetricX (MX) ↓	**2.8019**	2.8158
XCOMET (XC) ↑	**0.8738**	0.8730
chrF++ ↑	64.46	**64.69**
TER ↓	46.33	**46.25**
BLEU ↑	31.54	**31.76**
Environmental Impact		
Duration (min) ↓	753.79	**10.24**
Emissions (g) ↓	589.88	**6.39**
Total Energy (kWh) ↓	3.39	**0.037**
GPU Energy (kWh) ↓	2.5119	**0.0359**

Table 3: GP and GPO efficiency comparison.

the GP run suggests a substantial bottleneck inherent to the complexity of the original template rather than the inference engine itself. Consequently, this remains a valid comparison, as it highlights how template complexity can negate engine-level optimizations. Furthermore, Ollama lacks native support for resource-intensive Jinja2-based templates, necessitating streamlined adaptations for operational viability. Accordingly, all subsequent experiments were conducted using `Unsloth`, except for `4b_noft_gpo` variants.

To ensure computational efficiency and minimize the environmental footprint, this comparison was restricted to the ES-EN direction, excluding GP regardless of the engine. GPO was used in the remaining experiments, along with the previously mentioned minimal prompt. The absence of "GPO" in a system ID denotes the use of the minimal prompt.

4.2. Prompt, fine-tuning and direction

Regarding prompting, the performance of the GPO configuration is compared against the minimal prompt. As shown in Table 4, the performance gap between `4b_noft_gpo` and `4b_noft` is consistent across both translation directions, particularly in TER. In a zero-shot setting, the prompt serves as the sole mechanism to activate the model's translation capabilities. GPO more effectively triggers the model's internal weights by aligning with the linguistic distribution encountered during its primary training.

As for the fine-tuned models (indicated by the "ft" suffix), distinct patterns emerge depending on the translation direction. On the one hand, Table 4 shows that for ES-EN, performance is generally superior with the minimal prompt. On the other hand, it reveals the opposite trend for EN-ES, where GPO-based systems consistently outperform the minimal approach. In the EN-ES direction, GPO is superior as it aligns the model with its native instruction-tuning patterns (Finkelstein et al., 2026) and Spanish morphological requirements. Furthermore, using original, well drafted, Spanish annual reports as references during evaluation allows GPO to achieve native-level prose, effectively avoiding *translationese* (Zhang and Toral, 2019) and obtaining extremely high fluency scores in MetricX. Conversely, ES-EN finds benefit from using the minimal prompt. As demonstrated by Etxaniz et al. (2024), LLMs often fail to leverage their full potential when prompted in non-English languages, confirming an English-centric bias where less structural guidance in English leads to more natural generation. This is further supported by Richburg and Carpuat (2024), who found that the impact of translation fine-tuning is inherently uneven across language pairs. Finally, neural metrics like MetricX apparently exhibit a fluency bias, potentially over-rewarding the natural phrasing of the generated Spanish text, sometimes at the expense of strict lexical fidelity (Freitag et al., 2024).

Furthermore, reference quality is a primary factor explaining the discrepancy in MetricX scores across language directions, where EN-ES performance appears significantly superior to ES-EN (1.4895 vs. 2.5841). This may seem counterintuitive, as the source corpus is originally Spanish. As noted in the WMT23 findings: "Metrics might be guilty, but references are not innocent" (Freitag et al., 2023). By comparing the REFQA scores of the references with the QE scores of the model predictions, a clear limitation can be identified. If a metric deems a reference poorly aligned with the source, fine-tuning the model to mimic that reference may propagate those perceived flaws. This is evident when comparing fine-tuned results to the 4b_noft_GPO zero-shot scores. In the ES-EN direction, the MetricX gap between REFQA (2.9171) and QE (2.3320) indicates that the model's independent translations outperform the human references according to the metric. The same trend is observed in EN-ES, though absolute error scores are lower (REFQA 1.9424 vs. QE 1.6859). This proportional difference confirms that the EN-ES direction offers more room for improvement in terms of the metric.

4.3. Size of the model

Table 4 suggests a small difference between the 4b and the 12b variants (`4b_ft_bs_mix` vs. `12b_ft_bs_mix`). These two configurations were almost identical except that the 12b variant was loaded in 4-bit precision for inference and required minor tweaks due to VRAM constraints. The 12b variant is slightly superior, but given the increased resource consumption, the 4b variant appears to be a better option for deployment.

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
			ES-EN				
4b_noft_gpo	**2.5841**	**2.3320**	0.9300	**0.9398**	64.69	46.25	31.76
12b_ft_bs_mix	2.6666	2.7468	**0.9342**	0.9392	**70.73**	**38.05**	**40.15**
4b_ft_bs_mix	2.8291	2.8407	0.9253	0.9350	70.19	39.37	39.14
4b_ft_mix	2.8915	2.9337	0.9208	0.9305	69.20	40.51	37.80
4b_noft	2.9153	2.6358	0.9027	0.9139	62.99	223.79	32.16
4b_ft_gpo_bs_mix	2.9734	3.0378	0.9262	0.9362	69.63	39.84	38.51
4b_ft_gpo_mix	3.0343	3.0972	0.9206	0.9286	68.54	41.27	37.22
4b_ft_bs_ori	3.1675	3.4858	0.8885	0.8985	67.21	42.23	36.63
4b_ft_ori	3.2863	3.6783	0.8653	0.8742	65.68	43.72	35.28
4b_ft_gpo_ori	3.3970	3.8776	0.9169	0.9325	65.96	45.44	34.86
4b_ft_gpo_bs_ori	3.8412	4.7432	0.9176	0.9345	64.73	47.91	33.85
REFQA	–	2.9171	–	0.9153	N/A	N/A	N/A
			EN-ES				
12b_ft_gpo_bs_mix	**1.4895**	**1.6859**	**0.9524**	**0.9453**	**72.07**	**36.48**	**43.84**
4b_noft_gpo	1.5761	1.4672	0.9506	0.9544	64.80	47.99	34.24
4b_ft_gpo_bs_mix	1.8024	2.0062	0.9476	0.9429	69.65	40.10	40.99
4b_ft_gpo_mix	1.8058	1.9191	0.9391	0.9345	68.44	41.54	39.35
4b_ft_gpo_ori	1.8823	1.9493	0.9239	0.9231	67.30	42.78	38.35
4b_ft_gpo_bs_ori	1.9665	2.2565	0.9302	0.9305	68.04	41.21	39.79
4b_noft	1.9716	1.8419	0.9122	0.9164	63.24	403.09	34.87
4b_ft_mix	2.0267	2.1913	0.9399	0.9374	67.72	42.41	38.99
4b_ft_bs_mix	2.2443	2.5641	0.9464	0.9432	68.10	41.55	39.78
4b_ft_bs_ori	2.3601	2.9595	0.8261	0.8273	64.28	44.07	36.84
4b_ft_ori	2.5394	3.0886	0.8317	0.8242	62.52	46.51	35.05
REFQA	–	1.9424	–	0.9131	N/A	N/A	N/A

Table 4: Scores sorted by MetricX. Averaged from 1169 samples.

4.4. Variable-sized context results

In order to assess how the variable-sized context strategies (denoted as "mix") compare to regular fine-tuning, the same test set comprising 1,169 samples was split into two groups: one group with contexts (ES + EN pairs) of 1,000 tokens or more, and the other group with contexts of less than 1,000 tokens. This comparison is provided in Tables 5 and 6.

On the one hand, in both language directions, the short-context group performed similarly to those in the default table, with a slight improvement, mainly in MetricX. Since the difference in the number of samples is small (1,121 vs. 1,169), this behavior is expected. On the other hand, the scores of the long-context group (48 samples) are worth discussing. First of all, the biggest difference lies in MetricX, with scores significantly worse than those of the short-context group. In fact, the application of MetricX to extended contexts suggests a significant technical limitation regarding the cumulative nature of its scoring mechanism. As a regression-based model that evaluates a document as a single holistic unit, MetricX tends to aggregate minor stylistic and terminological deviations from the human reference across the entire text. While these discrepancies might be negligible in shorter segments, they apparently accumulate into an artificially inflated error score in long contexts, as the model lacks a mechanism to distinguish between a single catastrophic error and a series of consistent but technically acceptable stylistic variances. Furthermore, the model's underlying calibration is largely derived from human-annotated datasets such as MQM, which are predominantly composed of sentence-level or short-paragraph fragments. Consequently, the metric's regression head is optimized for short-span judgments and may not scale linearly or accurately when forced to process long contexts.

As for XCOMET scores, given the segmentation approach used to fit the 512-token limit, this metric can potentially fall short of assessing a larger context as a whole. Overall, it provides consistent scores across both groups, with a small decrease in the long-context group. Similarly, traditional metrics suggest good performance in the long-context group, on par with the short-context group, except for the lower half of the ranking, where the original dataset (`ori`) without GPO obtains poor scores, especially in the EN-ES direction.

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
Long contexts with ≥ 1,000 tokens (48 samples)							
12b_ft_bs_mix	**8.0723**	5.7770	0.8916	0.9053	**72.35**	**42.36**	**46.93**
4b_noft_gpo	8.1615	**5.6664**	0.9313	0.9263	68.96	46.66	41.09
4b_noft	8.3184	5.8418	0.8820	0.9117	68.70	47.54	40.93
4b_ft_gpo_bs_ori	8.4941	6.2150	0.9013	0.9281	68.11	51.95	41.04
4b_ft_gpo_bs_mix	8.5254	5.9954	0.9131	0.9265	71.63	45.44	45.08
4b_ft_bs_mix	8.5664	6.0687	0.9238	**0.9306**	71.75	46.06	44.90
4b_ft_mix	8.6068	6.1364	0.9020	0.9219	71.37	45.52	44.92
4b_ft_gpo_mix	8.6895	6.2077	0.9032	0.9210	70.65	46.28	44.06
4b_ft_gpo_ori	8.7376	6.4365	0.8951	0.9144	67.29	52.17	40.71
4b_ft_bs_ori	8.7539	7.6302	**0.9425**	0.9100	60.34	57.67	35.90
4b_ft_ori	9.2630	8.7100	0.8743	0.9181	55.17	62.29	32.49
Short contexts with ≤ 1,000 tokens (1121 samples)							
4b_noft_gpo	**2.3453**	**2.1892**	0.9300	0.9404	64.51	46.23	31.36
12b_ft_bs_mix	2.4351	2.6170	**0.9360**	**0.9407**	**70.66**	**37.86**	**39.86**
4b_ft_bs_mix	2.5835	2.7025	0.9254	0.9352	70.13	39.09	38.89
4b_ft_mix	2.6468	2.7965	0.9216	0.9308	69.10	40.30	37.50
4b_noft	2.6839	2.4985	0.9036	0.9140	62.74	231.34	31.79
4b_ft_gpo_bs_mix	2.7356	2.9112	0.9267	0.9366	69.55	39.60	38.23
4b_ft_gpo_mix	2.7921	2.9640	0.9213	0.9289	68.45	41.05	36.93
4b_ft_bs_ori	2.9283	3.3083	0.8862	0.8980	67.51	41.57	36.66
4b_ft_ori	3.0304	3.4628	0.8649	0.8723	66.13	42.93	35.39
4b_ft_gpo_ori	3.1683	3.7681	0.9179	0.9332	65.90	45.16	34.61
4b_ft_gpo_bs_ori	3.6419	4.6802	0.9183	0.9347	64.59	47.74	33.55

Table 5: Ablation analysis of the **ES-EN** pair size impact. Sorted by MetricX.

Regardless of these metric limitations, fine-tuning seems to improve traditional metric scores, with a bigger advantage in the short-context group compared to zero-shot. The `mix` runs using the variable-sized strategies are generally better than their `ori` counterparts in both short and long contexts. However, zero-shot `4b_noft_gpo` offers surprisingly good adaptability to variable-sized contexts off-the-shelf, ranking better than most of the fine-tuned systems in neural metrics, only behind the 12B systems in several metrics, especially the traditional ones.

4.5. Qualitative insights

Regarding `mix` runs, an inspection of the examples with the highest error scores ($MetricX > 10.0$) revealed no evident mistakes made by the models. This further suggests that MetricX has limitations when evaluating long-form contexts.

As for `4b_noft`'s noticeable poor performance in TER, in many instances, this flagging of low-quality segments was attributable to the production of an over-explanation and/or the creation of multiple alternative renderings, causing its output to be heavily penalized by TER's scoring mechanism. This highlights that neural models fail to penalize this kind of typical LLM production, probably exhibiting metric bias or reward hacking. In contrast, traditional metrics, mainly TER, were able to detect the deviation. This configuration with neither fine-tuning nor GPO would also produce its prediction in the source language, consequently yielding extremely poor scores in every metric.

It is worth noting the considerable score disparity observed in translations that received very low ratings despite being arguably correct in fine-tuned systems. One plausible explanation lies in the penalization of terminology errors arising from the translation of product names or institutional references. Most banks and financial institutions operate with established in-house terminology and defined translation strategies (for instance, a no-translation policy). Consequently, a fine-tuned system aligned to these stylistic choices is prone to deviate substantially from the metric criteria, since metrics do not consider these institutional style guidelines. Furthermore, professional human translators often adopt a functionalist approach, prioritizing dynamic equivalence Nida (1964, p. 159) and strategic localization over literal mapping. While these human-centric adaptations ensure the text is "fit for purpose" in a corporate context, they are frequently penalized by automated metrics that favor semantic overlap and stylistic uniformity over

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
Long contexts with ≥ 1,000 tokens (48 samples)							
12b_ft_gpo_bs_mix	**5.9720**	5.1175	0.9161	0.9252	**73.94**	**40.04**	**50.07**
4b_ft_gpo_bs_mix	6.3643	5.1175	0.8980	0.9232	70.44	46.31	45.48
4b_noft_gpo	6.3792	**4.7002**	**0.9455**	**0.9413**	69.79	47.22	42.54
4b_ft_gpo_mix	6.4082	5.2314	0.8821	0.8958	71.11	46.04	45.49
4b_noft	6.4538	4.7402	0.8954	0.8709	70.22	46.53	43.27
4b_ft_mix	7.1188	5.4089	0.9243	0.9170	65.80	55.37	40.39
4b_ft_bs_mix	7.4414	5.6484	0.9317	0.9308	60.15	58.21	35.67
4b_ft_gpo_ori	8.2959	6.5846	0.9226	0.9326	52.66	65.89	29.83
4b_ft_gpo_bs_ori	8.3613	7.7751	0.9350	0.9148	49.68	67.07	28.74
4b_ft_bs_ori	10.2780	13.6107	0.8307	0.8333	19.86	89.40	7.58
4b_ft_ori	10.9902	14.1523	0.8491	0.7497	18.70	89.53	7.79
Short contexts with ≤ 1,000 tokens (1121 samples)							
12b_ft_gpo_bs_mix	**1.2976**	1.5390	**0.9540**	0.9462	**71.99**	**36.33**	**43.57**
4b_noft_gpo	1.3704	**1.3287**	0.9508	**0.9550**	64.59	48.03	33.89
4b_ft_gpo_bs_mix	1.6071	1.8729	0.9497	0.9437	69.61	39.83	40.80
4b_ft_gpo_ori	1.6077	1.7508	0.9239	0.9227	67.92	41.79	38.72
4b_ft_gpo_mix	1.6087	1.7773	0.9416	0.9362	68.33	41.35	39.09
4b_ft_gpo_bs_ori	1.6927	2.0202	0.9300	0.9312	68.82	40.10	40.26
4b_noft	1.7797	1.7178	0.9129	0.9184	62.94	418.36	34.51
4b_ft_mix	1.8086	2.0535	0.9406	0.9383	67.80	41.86	38.93
4b_ft_bs_ori	2.0211	2.5034	0.8259	0.8270	66.18	42.13	38.09
4b_ft_bs_mix	2.0218	2.4321	0.9470	0.9437	68.44	40.83	39.96
4b_ft_ori	2.1776	2.6149	0.8309	0.8274	64.40	44.67	36.22

Table 6: Ablation analysis of the **EN-ES** pair size impact. Sorted by MetricX.

nuanced, contextual translation. These factors further explain why REFQA scores may underperform relative to QE ones.

4.6. Comparing results to other works

No direct comparisons with prior work were identified. The most closely related research is discussed below. For instance, the WMT24 repository[2] employed `google/metricx-23-xl-v2p0` (Juraska et al., 2023), an earlier version than the one utilized in this study, and `Unbabel/wmt23-cometkiwi-da-xl`, a reference-free metric comparable to XCOMET_QE scores. The winning system of that edition, Unbabel-Tower70B (Rei et al., 2024), achieved a MetricX score of 1.875 and a CometKiwi score of 0.745. At the time of writing, the WMT25 systems repository had not yet been released.

This prior work (Rajaee et al., 2026) reports XCOMET-XL scores for several language pairs, excluding Spanish, with averages ranging from 74.9 to 80.4. Another study (Oncevay et al., 2025) provides chrF++ and COMET scores for various language pairs; for English-source combinations, chrF++ results range from 43.79 to 66.8, while COMET scores range from 73.06 to 90.87. However, these findings are not directly comparable to the experiments in this study.

[2]https://github.com/wmt-conference/wmt24-news-systems

5. Conclusions

Addressing the research questions: (1) Despite TranslateGemma's notable zero-shot performance, fine-tuning on annual reports better aligns predictions with the reference style—as indicated by traditional metrics and qualitative analysis—even when neural metrics exhibit lower alignment with human references. (2) IBEX 35 companies may find value in this approach given the remarkably low TER, which minimizes Machine Translation Post-Editing (MTPE) effort. (3) The variable-sized context strategies for fine-tuning provide slightly better results in both short- and long-context settings. (4) There seem to be limitations in current neural MT metrics, including metric bias toward model predictions, reward hacking effects, and language direction asymmetry. However, a thorough qualitative analysis should be performed on a larger scale in order to empirically affirm whether that is the reason for the REFQA underperformance against zero-shot QE, consequently limiting fine-tuning effectiveness.

6. Acknowledgements

This work is framed under the Spanish National Project GRESEL (PID2023-151280OB-C21). It was also partially funded by grant PTA2023-023812-I (awarded to Yanco Amor Torterolo Orta) through MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+); and by an FPI-UAM scholarship awarded to Melina Chatzi.

7. Bibliographical References

Emad A. Alghamdi, Jezia Zakraoui, and Fares A. Abanmy. 2023. Domain adaptation for arabic machine translation: The case of financial texts.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages dialects.

Paolo Di Natale, Elena Chiocchetti, and Egon Waldemar Stemle. 2025. Meta-evaluation of automatic machine translation metrics between Italian and a minor language variety of German. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 371–383, Cagliari, Italy. CEUR Workshop Proceedings.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. Translategemma technical report.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023. On search strategies for document-level neural machine translation.

Leticia Herrero Rodes and Verónica Román Mínguez. 2015. English to spanish translation of the economics and finance genres. *InTRAlinea: Online Translation Journal*, (Special Issue: New Insights into Specialised Translation). Revista del Departamento de Traducción e Interpretación de la Universidad de Bolonia.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Juraj Juraska. 2025. Comment on metricx-25 (issue #12). GitHub Issue Comment. Google Research MetricX Repository.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968,

Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31033–31059, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Haijun Li, Tianqi Shi, Zifu Shang, Yuxuan Han, Xueyu Zhao, Hao Wang, Yu Qian, Zhiqiang Qian, Linlong Xu, Minghao Wu, Chenyang Lyu, Longyue Wang, Gongbo Tang, Weihua Luo, Zhao Xu, and Kaifu Zhang. 2025. Transbench: Benchmarking machine translation for industrial-scale applications.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. Dolfin – document-level financial test set for machine translation.

Eugene A. Nida. 1964. *Toward a Science of Translating*. E. J. Brill, Leiden.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Albuquerque, New Mexico. Association for Computational Linguistics.

Cristóbal Parra Quesada and Manuela Cañizares Espada. 2024. Relationship of market capitalization of the ibex 35 to corporate social responsibility and transparency. *Corporate Social Responsibility and Environmental Management*, 31(4):3551–3572.

Sara Rajaee, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. 2026. Unlocking reasoning capability on machine translation in large language models.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages

185–204, Miami, Florida, USA. Association for Computational Linguistics.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation?

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

8. Language Resource References

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Sofía Micaela Roseti, Blanca Carbajo-Coronado, and Jordi Porta. 2025. Financial ES-EN parallel corpus from annual reports.

LabelFusion: Fusing Large Language Models with Transformer Encoders for Robust Financial News Classification

Michael Schlee[1], **Christoph Weisser**[2], **Timo Kivimäki**[3]
Melchizedek Mashiku[4], **Benjamin Säfken**[5]
[1]Centre for Statistics, Georg-August-Universität Göttingen, Germany
[2]Hochschule Bielefeld (HSBI) - University of Applied Sciences and Arts, Bielefeld, Germany
[3]Department of Politics and International Studies, University of Bath, Bath, UK
[4]Tanaq Management Services LLC, Contracting Agency to the Division of Viral Diseases
Centers for Disease Control and Prevention, Chamblee, Georgia, USA
[5]Institute of Mathematics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany
michael.schlee@uni-goettingen.de, christoph.weisser@hsbi.de, t.kivimaki@bath.ac.uk
melchizedek.mashiku@tanaq.com, benjamin.saefken@tu-clausthal.de

Abstract

Financial news plays a central role in shaping investor sentiment and short-term dynamics in commodity markets. Many downstream financial applications—such as commodity price prediction or sentiment modeling—therefore rely on the ability to automatically identify news articles that are relevant to specific assets. However, obtaining large labeled corpora for financial text classification tasks is costly, and transformer-based classifiers such as RoBERTa often degrade significantly in low-data regimes. Our results show that appropriately prompted out-of-the-box Large Language Models (LLMs) achieve strong performance even in low-data regimes. Furthermore, we propose LabelFusion, a hybrid architecture that combines the output of a prompt-engineered LLM with contextual embeddings produced by a fine-tuned RoBERTa encoder through a lightweight Multilayer Perceptron (MLP) voting layer. Evaluated on a ten-class multi-label subset of the Reuters-21578 corpus, LabelFusion achieves a macro F1 score of 96.0% and an accuracy of 92.3% when trained on the full dataset, outperforming both standalone RoBERTa (F1 94.6%) and the standalone LLM (F1 93.9%). In low- to mid-data regimes, however, the LLM alone proves surprisingly competitive, achieving an F1 score of 75.9% even in a zero-shot setting and consistently outperforming LabelFusion until approximately 80% of the training data is available. These results suggest that LLM-only prompting represents the preferred strategy under annotation constraints, whereas LabelFusion becomes the most effective solution once sufficient labeled data is available to train the encoder component. The code is available in an anonymized repository.

Keywords: multi-label text classification, large language models, fusion model, financial NLP, Reuters-21578

1. Introduction

Emotional factors play an substantial role in the decision-making processes of both institutional and individual investors. Such emotional signals exert a measurable influence on individual commodity returns, including assets such as oil or gold, and can even enable short-term predictive insights (Sinha and Shastri, 2020).

News articles on specific commodities can act as a primary source of these emotional signals; consequently, financial news represents a valuable resource for identifying market sentiment and potentially anticipating subsequent developments in commodity prices.

Transformer-based models such as FinBERT are commonly employed to extract sentiment from financial news by fine-tuning pre-trained language models on labeled financial text corpora (Araci, 2019). Different model families integrate these sentiment scores in different ways. Hybrid Long Short-Term Memory (LSTM) models use transformer-extracted sentiment as additional input features that vary in time along with price data (Chae and Choi, 2023; Yang et al., 2022; Nabipour et al., 2026). Transformer-based approaches incorporate sentiment through attention mechanisms that weight emotional signals according to their relevance to price movements (Chen et al., 2024).

An important preliminary step in news-based financial prediction tasks, such as forecasting sentiment in news to predict future gold or oil prices, is the filtering of relevant articles. Only news that is directly related to the specific commodity provides meaningful and informative training data for subsequent modeling tasks. This data filtering step is therefore as critical as the sentiment classification task itself. The extremely large volume of available financial news makes manual selection infeasible, creating a clear need for automated text classification methods capable of identifying documents that are relevant to specific commodities or financial assets.

LabelFusion addresses this gap by combining state-of-the-art LLMs with a fine-tuned RoBERTa encoder through an intelligent MLP-based voting mechanism. This approach enables effective multi-label text classification while reducing dependence on extensive manually labeled training data. Our contributions are as follows:

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 98–105

1. We show that out-of-the-box LLMs, when used with appropriate prompting, achieve high classification accuracy even when training data is scarce.

2. We introduce a hybrid fusion model that combines LLM predictions with RoBERTa embeddings via a trainable MLP. With sufficient training data (80%–100%), the approach improves the F1 score from 0.939 to 0.960 compared to a standalone fine-tuned RoBERTa model.

3. We present *LabelFusion*, a user-friendly software package that facilitates the integration and deployment of fusion-based classification models (Anonymous, 2025).

2. Related Work

The Reuters-21578 corpus (Lewis, 1997) has served as the standard benchmark for multi-label financial news categorization for decades, with classical classifiers such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Rocchio establishing strong baselines under severe label imbalance (Debole and Sebastiani, 2005). More recent work has applied Graph Convolutional Networks (GCNs) that propagate semantic information across document, word, and label nodes (Zeng et al., 2024), as well as attention-based architectures that explicitly model label correlations (Yuan et al., 2024; Ma et al., 2024). Despite their strong empirical results, all of these methods depend on substantial labeled training data and do not incorporate the broad language understanding that LLMs provide.

Brown et al. (2020) demonstrated that GPT-3 achieves competitive performance across a wide range of Natural Language Processing (NLP) benchmarks through in-context few-shot prompting without any gradient updates, an intuition formalized by Schick and Schütze (2021) through cloze-style Pattern-Exploiting Training (PET) and refined by Gao et al. (2021) via automatic prompt and verbalizer search. More recent evaluations confirm that instruction-tuned LLMs are effective zero-shot classifiers (Wang et al., 2023), although fine-tuned encoder models retain a competitive advantage when sufficient labeled data is available (Chae and Davidson, 2025).

Ensemble and fusion approaches that combine predictions from multiple pre-trained models consistently outperform individual classifiers (Abburi et al., 2023), and more tightly integrated architectures show that encoder embeddings and LLM-derived features contribute complementary information (Koloski et al., 2024; Gwak and Jung, 2025). To our knowledge, no prior work has explicitly fused prompt-based LLM predictions with fine-tuned encoder representations for multi-label financial news classification. LabelFusion addresses this gap by combining both sources through a trainable MLP voting layer, enabling robust performance across the full spectrum of labeled data availability.

3. Model Architecture

Let $\mathbf{x}$ denote an input text to be assigned labels from a predefined label set $\mathcal{Y} = \{1, \ldots, K\}$, where K denotes the total number of possible categories. The task is multi-label classification, meaning that multiple labels may be associated with a single input text. LabelFusion combines two complementary components — a prompt-based LLM and a fine-tuned RoBERTa encoder — whose outputs are fused by a trainable MLP voting layer.

3.1. Prompt-Based LLM Component

The input text $\mathbf{x}$ is inserted into a prompt template

$$p(\mathbf{x}) = \mathcal{T}(\mathbf{x}, \mathcal{Y}, \mathcal{E}),$$

where $\mathcal{T}(\cdot)$ denotes the prompt template, $\mathcal{Y}$ represents the list of predefined labels, and $\mathcal{E}$ denotes an optional set of demonstration examples. Three prompt regimes are considered: *zero-shot* ($\mathcal{E} = \emptyset$), *one-shot* ($|\mathcal{E}| = 1$), and *few-shot* ($|\mathcal{E}| > 1$). The prompt is processed by a LLM

$$f_{\text{LLM}} : p(\mathbf{x}) \mapsto \mathbf{z},$$

where $\mathbf{z} \in \{0, 1\}^K$ is a binary prediction vector. Each element z_k is defined as $z_k = 1$ if label k is predicted to be present and $z_k = 0$ otherwise.

3.2. RoBERTa Representation Component

The same input text is independently encoded by a fine-tuned RoBERTa model:

$$h = f_{\text{RB}}(\mathbf{x}) \in \mathbb{R}^{768},$$

where h corresponds to the contextual embedding of the `[CLS]` token, capturing rich task-specific semantic information from the input text.

3.3. Feature Fusion and Prediction

The outputs of both components are concatenated to form a fused representation:

$$u = [h\,;\, z] \in \mathbb{R}^{768+K}.$$

This fused vector combines the dense contextual embedding from RoBERTa with the discrete label

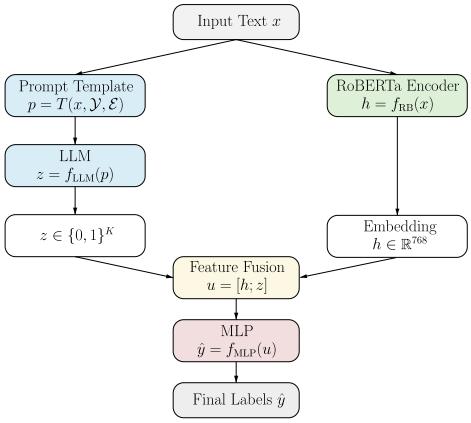


Figure 1: Architecture of *LabelFusion*. An Input text x is processed along two parallel branches: a prompt template $T(x, \mathcal{Y}, \mathcal{E})$ feeds a LLM to produce a binary label vector $z \in \{0,1\}^K$, while a RoBERTa encoder produces a contextual embedding $h \in \mathbb{R}^{768}$. The two representations are concatenated into a joint feature vector $u = [h\,;\, z]$, which is passed to a trainable MLP that outputs the final label predictions $\hat{y}$.

predictions from the LLM, and is passed through a MLP voting model:

$$\hat{\mathbf{y}} = f_{\text{MLP}}(u) \in [0,1]^K,$$

where each element $\hat{y}_k$ represents the predicted probability of label k. The architecture is illustrated in Figure 1.

4. Experimental Setup

4.1. Dataset

The dataset was constructed from the Reuters-21578 (R21578) corpus, which contains approximately 12,902 documents spanning around 135 topic categories with a highly skewed label distribution. To obtain a manageable and well-populated benchmark, only the ten most frequent topics were retained: *earn*, *acq*, *money-fx*, *grain*, *crude*, *trade*, *interest*, *ship*, *wheat*, and *corn*. Documents containing none of the selected topics were excluded, yielding 9,034 documents. Each article was represented using its full raw text, and a multi-label binary target vector of length ten was created for each document. Despite the filtering, the dataset remains challenging: label frequencies range from 3,964 documents for *earn* to only 237 for *corn*, and 9.2% of documents carry two or more topic labels simultaneously.

The corpus provides an official benchmark split into training and test data, which was preserved to ensure comparability with prior work. The test set (2,545 documents) was kept intact from the original ModApte split. From the training portion, a validation set of 647 documents (10%) was carved out using multilabel stratified sampling to maintain a consistent topic distribution across splits, leaving 5,842 documents for training. This procedure yielded three subsets: a training set used for model learning, a validation set for hyperparameter tuning and model selection, and a held-out test set used exclusively for the final evaluation of model performance.

4.2. Experimental Setup

To evaluate the proposed LabelFusion architecture, we conduct experiments with varying amounts of labeled training data. We construct training subsets comprising 5%, 10%, 20%, 40%, 60%, 80%, and 100% of the available labeled dataset (5,842 samples). Each subset is used to train the supervised components of the model (RoBERTa and the fusion MLP), while the LLM branch remains prompt-based. For each subset, the respective proportion of training data is incorporated into the prompt template, which is then sent to the LLM.

We compare LabelFusion against several baselines:

- RoBERTa, a transformer encoder used as a standalone classifier, is fine-tuned on the respective proportion of the training data.
- GPT-5-nano receives training data subsets embedded in our prompt template. Additionally, we evaluate the performance of the out-of-the-box model in a zero-shot setting, where no training data is used and the prompt template is not applied. Furthermore, we assess classification performance in ultra-low data scenarios, where only a single example is included in the prompt template.
- Term Frequency–Inverse Document Frequency (TFIDF) + Logistic Regression (LR), a classical linear baseline trained on the full dataset.

For each configuration, we report Accuracy, F1 score, Precision, and Recall. These metrics allow us to evaluate both overall prediction quality and the balance between false positives and false negatives.

5. Results & Discussion

Table 5 reports macro F1, accuracy, precision, and recall across all training data fractions and models. The classical TFIDF+LR baseline achieves 80.6%

Data	Model	F1	Acc.	Prec.	Rec.
0-shot	GPT-5-nano	75.9	83.4	89.2	70.5
1-shot	GPT-5-nano	**76.1**	84.1	92.0	68.0
5% (292)	Fusion	71.7	70.6	72.0	71.5
5% (292)	RoBERTa	37.2	0.0	27.6	71.3
5% (292)	GPT-5-nano	**93.0**	88.1	95.2	91.7
10% (584)	Fusion	67.1	67.0	67.2	67.1
10% (584)	RoBERTa	41.7	40.0	32.1	61.6
10% (584)	GPT-5-nano	**93.8**	88.5	96.2	92.6
20% (1168)	Fusion	75.2	72.0	76.9	74.5
20% (1168)	RoBERTa	53.4	67.3	46.5	64.3
20% (1168)	GPT-5-nano	**92.8**	88.6	95.1	92.3
40% (2336)	Fusion	88.6	83.6	89.3	88.9
40% (2336)	RoBERTa	83.6	82.0	85.8	85.0
40% (2336)	GPT-5-nano	**93.1**	87.9	95.2	91.7
60% (3505)	Fusion	93.2	85.5	92.9	95.0
60% (3505)	RoBERTa	90.7	83.4	90.6	94.5
60% (3505)	GPT-5-nano	**93.8**	88.4	95.9	92.4
80% (4673)	Fusion	**95.4**	90.2	95.4	96.5
80% (4673)	RoBERTa	94.3	88.8	93.0	96.6
80% (4673)	GPT-5-nano	93.4	88.0	95.1	91.8
100% (5842)	Fusion	**96.0**	92.3	96.7	96.1
100% (5842)	RoBERTa	94.6	89.0	93.2	96.6
100% (5842)	GPT-5-nano	93.9	88.9	96.3	92.7
100% (5842)	TFIDF+LR	68.2	80.6	95.4	56.9

Table 1: Performance comparison of LabelFusion, standalone RoBERTa, GPT-5-nano, and a TFIDF + LR baseline across varying amounts of labeled training data. GPT-5-nano is evaluated in one-shot mode throughout; the zero-shot row reports its performance without any training data. TFIDF + LR is trained on the full dataset and included as a classical reference. All metrics are macro-averaged and reported in percent (%) on the Reuters-21578 test set.

accuracy and a macro F1 of 68.2%, with high precision (95.4%) but low recall (56.9%), reflecting conservative prediction behavior.

Several clear patterns emerge from Table 5. Even in zero-shot mode, the standalone LLM already achieves a macro F1 of 75.9%, exceeding the TFIDF+LR baseline (68.2%) by a large margin, likely due to the broad general knowledge and natural language understanding acquired during pretraining. Moving to one-shot mode yields only a marginal gain (76.1%), suggesting that the LLM requires more than a single demonstration to benefit meaningfully from in-context examples.

In the ultra-low data regime, the standalone RoBERTa classifier struggles severely. With only 5% of the training data (292 documents), it achieves an accuracy of 0.0% and a macro F1 score of 37.2%, indicating that the model has not yet learned a reliable decision boundary. With 10% of the data (584 documents), performance improves to an F1 score of 41.7%, but remains far below the LLM. In contrast, the LLM achieves an F1 score of 93.0% at 5% and 93.8% at 10%, demonstrating that it is a highly effective standalone solution in low-data settings. LabelFusion in these regimes achieves F1 scores of 71.7% and 67.1% at 5% and 10%, which are better than standalone RoBERTa but inferior to the LLM alone. We interpret this as a consequence of the poorly trained RoBERTa component, which appears to introduce noise into the voting process and thereby degrades the overall fusion performance relative to the LLM in isolation.

In the low- to mid-data regime spanning 5% to 60% of the training data, the LLM consistently outperforms both standalone RoBERTa and LabelFusion with respect to macro F1 score. We further observe that using only 5% of the training data is sufficient for the LLM to achieve one of the highest F1 scores in the overall classification task. This finding highlights the potential of LLMs as a standalone approach. In the low- to mid-data regime, LabelFusion is still hampered by noise introduced by the insufficiently fine-tuned RoBERTa encoder, whose weak task-specific signal interferes with the more reliable LLM predictions in the voting layer. Nevertheless, a clear upward trend can be observed: as the proportion of training data increases, the RoBERTa component begins to contribute increasingly useful representations, leading to a steady improvement in LabelFusion's overall accuracy.

From 80% of the training data onward, this trend reaches a tipping point and the situation reverses decisively. LabelFusion surpasses both standalone methods, reaching an accuracy of 90.2% and a macro F1 score of 95.4% at 80%, and achieving the best overall performance with 92.3% accuracy and a macro F1 score of 96.0% when trained on the full dataset. These results confirm that once the RoBERTa component is sufficiently trained, the fusion mechanism effectively combines the task-specific discriminative power of the transformer with the broad language understanding of the LLM, resulting in a model that is more robust than either component in isolation.

6. Conclusion & Future Work

We introduced LabelFusion, a hybrid architecture that fuses the logits of a prompt-engineered LLM with embeddings from a fine-tuned RoBERTa encoder via a trainable MLP voting layer for multi-label financial news classification. Experiments on the Reuters-21578 corpus show that the two model families are complementary and that the optimal strategy depends on the available annotation budget: a carefully prompted LLM constitutes the strongest standalone solution in low- to mid-data regimes, while LabelFusion becomes the preferred choice once sufficient labeled data is available—approximately from 80% of the training data

onward—where the fine-tuned encoder provides reliable task-specific signals that the fusion layer can exploit effectively.

Future work will explore the integration of additional modalities into LabelFusion. The architecture is designed for the seamless integration of further input sources. In particular, we hypothesise that recent stock price time series of corresponding commodities significantly influence the probability that the commodity appears in financial news. The next logical step would therefore be the integration of historical commodity prices as an additional input modality. These time series could be processed by time-series transformer architectures, which are well suited for extracting short-term temporal patterns through attention mechanisms, and could thereby further improve the macro F1 score of the fusion model.

7. Acknowledgments

The work presented in this paper was conducted independently by the author Melchizedek Mashiku and is not affiliated with Tanaq Management Services LLC, Contracting Agency to the Division of Viral Diseases, Centers for Disease Control and Prevention, Chamblee, Georgia, USA.

8. Bibliographical References

Harika Abburi et al. 2023. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*.

Anonymous. 2025. Labelfusion: Learning to fuse llms and transformer classifiers for robust text classification.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models.

Tom B. Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Seung Chan Chae and Sun-Yong Choi. 2023. Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models. *Axioms*, 12(9):835.

Youngjin Chae and Thomas Davidson. 2025. Large language models for text classification. *Sociological Methods & Research*.

Yilin Chen, Xiaolong Li, and Yonghong Hu. 2024. Dual-attention transformer for financial news sentiment and stock price prediction. *Expert Systems with Applications*, 238:121134.

Franca Debole and Fabrizio Sebastiani. 2005. An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, 56(6):584–596.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *ACL*, pages 3816–3830.

Jungwoo Gwak and Yoonsang Jung. 2025. Layer-aware embedding fusion for llms. *arXiv preprint arXiv:2504.05764*.

Boshko Koloski, Senja Pollak, Roberto Navigli, and Blaž Škrlj. 2024. Automl-guided fusion of entity and llm-based representations. In *ICONIP*, pages 89–102.

David D. Lewis. 1997. Reuters-21578 text categorization test collection. Technical report, AT&T Labs Research.

Yilin Ma, Xin Gao, and Bowen Yang. 2024. LIAM: Label-interaction aware model for multi-label text classification. *Neurocomputing*, 580:127139.

Morteza Nabipour, Pejman Naeem, and Hamed Jabani. 2026. Federated learning for financial sentiment-augmented price prediction. In *Proceedings of the 2026 International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification. In *EACL*, pages 255–269.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 145–160, Berlin, Heidelberg. Springer.

Anand Sinha and Ravi Shastri. 2020. Impact of news sentiment on commodity returns. *Journal of Behavioral Finance*, 21(3):310–325.

Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, pages 22–35. PMLR.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Jianfeng Yang, Yufei Wang, and Xiang Li. 2022. Prediction of stock price direction using the LASSO-LSTM model combining technical indicators and financial sentiment analysis. *PeerJ Computer Science*, 8:e1148.

Hao Yuan, Weiguang Han, and Yucheng Li. 2024. LACN: Label-aware co-training network for multi-label text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 3412–3422. ELRA and ICCL.

Dingkun Zeng, Erxue Zha, Jian Kuang, and Yong Shen. 2024. Multi-label text classification based on semantic-sensitive graph convolutional network. *Knowledge-Based Systems*, 282:111107.

A. Dataset Construction and Processing

This section describes the construction of the dataset used in all experiments. We used the Reuters-21578 corpus as provided through the NLTK Reuters corpus interface and constructed a multi-label classification dataset based on the standard ModApte split.

A.1. Document Extraction

We restricted the label space to the ten most frequent Reuters topics commonly used in the ModApte setting: `earn`, `acq`, `money-fx`, `grain`, `crude`, `trade`, `interest`, `ship`, `wheat`, and `corn`.

Documents from both the original training and test partitions were filtered accordingly, and only those assigned at least one of the selected labels were retained in the final dataset.

For each retained document, the model input was constructed by concatenating the headline and the article body into a single text sequence.

A.2. Dataset Splitting Procedure

The original Reuters training and test partitions were preserved as defined by the corpus reader, where document identifiers beginning with `train` and `test` determine split membership. The filtered test partition was kept unchanged throughout all experiments as the held-out test set.

The filtered training partition was further divided into training and validation subsets using a fixed split ratio of 90/10. To preserve the multi-label distribution across the selected topic set, multilabel-stratified shuffle splitting based on the iterative stratification method was applied (Sechidis et al., 2011; Szymański and Kajdanowicz, 2017).

A.3. Dataset Assembly

All retained documents were aggregated into a tabular dataset in which each row corresponds to a single Reuters article. The final representation consisted of the document text and ten binary topic columns.

After filtering for the selected topic set, the dataset comprised 9,034 documents in total, including 5,842 training, 647 validation, and 2,545 test documents.

All preprocessing and split operations were deterministic to ensure reproducibility across experimental runs.

Topic	Training set	Validation set
earn	2877 (44.30%)	288 (44.51%)
acq	1650 (25.41%)	165 (25.50%)
money-fx	539 (8.30%)	54 (8.35%)
grain	434 (6.68%)	43 (6.65%)
crude	391 (6.02%)	39 (6.03%)
trade	369 (5.68%)	37 (5.72%)
interest	347 (5.34%)	35 (5.41%)
wheat	212 (3.26%)	21 (3.25%)
ship	198 (3.05%)	20 (3.09%)
corn	182 (2.80%)	18 (2.78%)

Table 2: Class distribution in the training and validation sets after multilabel-stratified splitting.

Parameter	Value
Base model	roberta-base
Maximum sequence length	256
Learning rate	2×10^{-5}
Number of epochs	2
Batch size	32
Task setting	Multi-label classification

Table 3: Hyperparameters of the standalone RoBERTa classifier.

A.4. Class Distribution

The distribution of topic labels in the training dataset reflects the natural class imbalance of the Reuters corpus. As shown in Table 2, the relative frequency of each topic remains nearly identical between the training and validation sets after multilabel-stratified splitting, demonstrating that the stratification procedure successfully preserved the overall class distribution across subsets.

B. Hyperparameter and Training Configuration

The configuration parameters for all model components and experimental settings are summarized in Tables 3–6. Unless stated otherwise, the same hyperparameters were applied across all experimental conditions.

B.1. Prediction Threshold

Predicted probabilities produced by the classification models were converted into binary label assignments using a fixed threshold of 0.5 for all topic categories. The same threshold was applied consistently across all models and experimental settings.

Parameter	Value
Model	GPT-5-nano
Temperature	0.1
Top-p	1.0
Maximum completion tokens	150
Task setting	Multi-label classification
Caching	Enabled

Table 4: Hyperparameters of the LLM classifier.

Parameter	Value
Fusion hidden dimensions	[64, 32]
Learning rate (RoBERTa in fusion)	1×10^{-5}
Learning rate (fusion MLP)	5×10^{-4}
Number of epochs	10
Batch size	8
Classification type	Multi-label

Table 5: Hyperparameters of the fusion ensemble model.

C. Prompt Template

For an input text x, the prompt presented to the language model follows a consistent structure consisting of task instructions, optional training examples, and an output format specification.

The initial role prompt is automatically generated from labeled training examples using a dedicated meta-prompting procedure. This design ensures that the prompt adapts to the characteristics of each dataset rather than relying on a fixed manually specified role description.

```
You are a financial analyst who
identifies whether a news article
belongs to one or more predefined
commodity-related categories.

The following examples show texts and
their classifications for labels:
earn, acq, money-fx, grain, crude,
trade, interest, ship, wheat, corn

Training Data Examples:
Example 1:
Text: U.S. grain exporters reported
increased shipments following strong
overseas demand.
Ratings: 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0
| 0 | 0

Example 2:
Text: The company announced plans to
acquire a regional competitor in a
stock transaction.
Ratings: 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0
| 0 | 0

Given the above information, make a
prediction for the following paragraph.

{test article}

The output format shall be:

    0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
0

Each value must be either 1 (label
present) or 0 (absent).
Multiple values may be 1, as this is a
multi-label classification task.
No additional text shall be shown --
output only the classification line.
```

Prompting regime	Number of examples
Zero-shot	0
One-shot	1
Few-shot	20

Table 6: Number of in-context examples used in each prompting regime.

LLM-as-a-Judge Evaluation of Financial News Articles generated based on Factors of Stock Price Fluctuation

Yurina Kosai, Yucheng Xie, Rikuto Tsuchida, Takehito Utsuro
Graduate School of Science and Technology, University of Tsukuba
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
{s2520763, s2620847, s2520796}_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

This paper proposes an LLM-as-a-Judge evaluation framework of stock price fluctuation articles automatically generated based on financial news, corporate disclosures, and stock price fluctuation data. This automatic article generation framework emulates the workflow of human financial journalists by analyzing recent stock price fluctuations and incorporating relevant causal factors extracted from textual and numerical information. In particular, the generation process utilizes news articles and numerical stock price data, including price fluctuation ranges over the past three days. Based on those automatically generated stock price fluctuation articles, this study places particular emphasis on the LLM-as-a-Judge evaluation methodology. We conduct an item wise human evaluation and compare it with the LLM-as-a-Judge automatic metric. We analyze the correlation among these evaluation methods to assess their reliability. Furthermore, through comparisons between zero-shot and few-shot prompting, we examine the effectiveness of the proposed framework and the validity of LLM based evaluation for assessing factual and causal consistency in financial text generation.

Keywords: LLM-as-a-Judge, automatic evaluation, generating stock price fluctuation articles, factors of stock price fluctuation, large language models

1. Introduction

In providing information on stock price fluctuations, the usefulness of news articles extends beyond merely reporting the magnitude of price changes. Such articles also offer insights into the underlying factors that have led to those fluctuations. Typically, these articles are manually written for each individual stock. The conventional writing procedure is assumed to follow the format illustrated in Figure 1. Specifically, for stocks exhibiting large price fluctuations (either increases or decreases), journalists first identify information that may have contributed to the observed fluctuations. Subsequently, for stocks where such information is found, journalists summarize the relevant content and compose articles based on it. To address this task, automatic generation of stock price fluctuation explanation articles using large language models (LLMs) has recently attracted attention. Existing automatic generation methods (Nishida and Utsuro, 2025) achieve this by referring to numerical information on stock price fluctuations together with textual information that may serve as potential causal factors. In these approaches, official corporate IR disclosures are collected as textual information related to stock price fluctuation factors and used as the primary information source.

In contrast, this paper focuses on stock price fluctuation explanation article generation methods using LLMs (Nishida and Utsuro, 2025) and proposes an automatic evaluation method for the generated articles based on the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023). Within the framework of this study, we first obtain stock price data for the most recent three days for each stock, based on the daily stock price change ranking published on the stock information website Kabutan[1], and use these data as factual information regarding stock price fluctuations[2]. Next, following (Nishida and Utsuro, 2025), we collect official corporate IR disclosures as textual information related to potential stock price fluctuation factors, and use these as the information source for generating stock price fluctuation explanation articles with an LLM. In addition, for these IR disclosures, we manually compose reference articles that explicitly clarify the relationship between the disclosed information and the stock price fluctuations. These serve as reference stock price fluctuation explanation articles. Based on these reference articles, we design a 10 point evaluation scale to assess content validity and clarity of explanation. Using this evaluation scale, we conduct both human evaluation of the automatically generated stock price fluctuation explanation articles and automatic evaluation under the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023). In the evaluation experiments, we apply the LLM-as-a-Judge automatic evaluation scale to stock price fluctua-

[1] https://kabutan.jp/

[2] It should be noted that the article texts published on Kabutan are not used at all; only numerical data are utilized.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 106–113

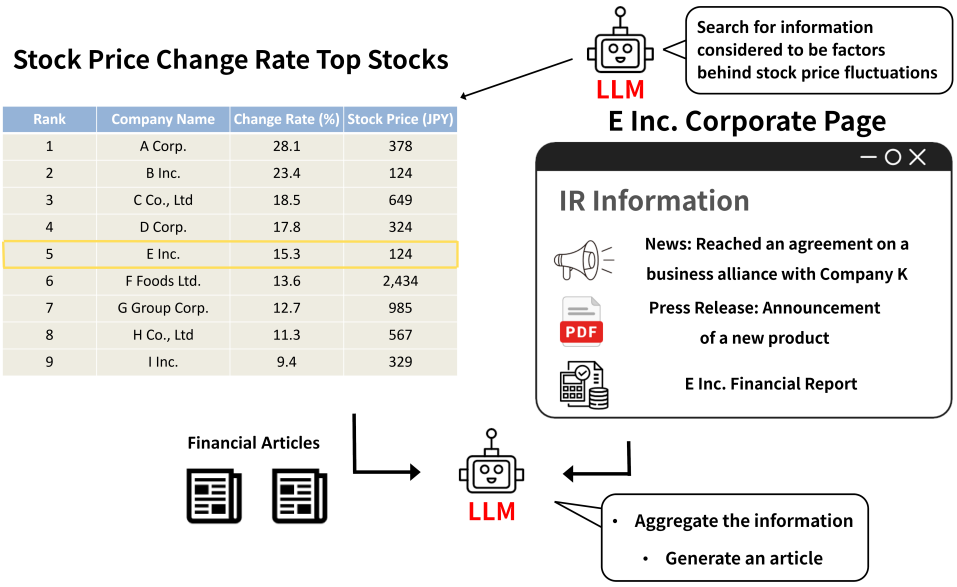


Figure 1: Generation of stock price fluctuation reason articles based on stock price fluctuation factor information (Nishida and Utsuro, 2025)

tion explanation articles generated by GPT-4o[3] in both zero-shot and few-shot settings, and achieve sufficiently high correlation with human evaluation results. In particular, higher correlation is observed for articles generated in the few-shot setting. In contrast, we show that the correlation between ROUGE scores (Lin, 2004) and human evaluation results is extremely low, thereby demonstrating the effectiveness of the proposed automatic evaluation scale under the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023).
The main contributions of this paper are summarized as follows:

- We propose an LLM based automatic evaluation framework for stock price fluctuation article generation, grounded in the LLM-as-a-Judge paradigm, to assess the semantic adequacy and explanatory quality of generated financial articles.
- We construct a benchmark setting that combines numerical stock price data and official corporate IR disclosures, while manually creating reference articles to ensure clear causal alignment between price fluctuations and explanatory content.
- We design a task specific 10 point evaluation scheme that decomposes stock price fluctuation articles into key components (stock price information, explanation of fluctuation factors, and supplementary information), allowing fine-grained comparison between human and automatic evaluations.
- Through experiments on GPT-4o generated articles (zero-shot and few-shot), we demonstrate that the proposed LLM based evaluation method achieves a strong correlation with human judgments, substantially outperforming ROUGE in this task.

2. Related Work

Regarding the LLM-as-a-Judge, which employs LLMs as evaluators, it has been reported that automatic evaluation by LLMs can serve as a viable alternative to human evaluation (Chiang and Lee, 2023; Zheng et al., 2023). For example, Zheng et al. (2023) demonstrated the potential of utilizing LLMs for assessing the subjective quality of generated responses. Furthermore, Liu et al. (2023) proposed a framework in which evaluation criteria are explicitly provided to the LLM, which then generate a chain-of-thought and perform step-by-step scoring in a form-based input format. Their results show that evaluation using GPT-4 substantially outperforms conventional automatic metrics in terms of correlation with human judgments. Meanwhile, Saha et al. (2024) proposed a branch-solve-merge approach, in which evaluation criteria are decomposed into multiple aspects, each of which is assessed individually, and the results are subsequently aggregated. They report that this method achieved higher agreement with human evaluation compared to single-pass holistic evaluation. In Imajo et al. (2025), a reference answer set based evaluation framework is constructed along three dimensions — fluency, truthfulness, and helpfulness — and its consistency with LLM based evaluation results is demonstrated.
In the research activities targeting question an-

[3] https://openai.com/ja-JP/api/

swering (QA) tasks where concrete reference answers to the questions do exist, Badshah and Sajjad (2025) proposed a majority voting evaluation method using multiple LLMs. They showed that combining multiple models improves evaluation reliability and achieved strong correlation with human evaluation. On the other hand, Bai et al. (2023) introduced stepwise scoring and ranking based on multiple criteria, including accuracy, coherence, factuality, and comprehensiveness, and reported that the resulting evaluation outcomes exhibit high agreement with human annotations.

3. Stock Price Fluctuation Factor Information

In this study, we use officially disclosed corporate information as potential factors underlying stock price fluctuations. Specifically, we collect IR disclosures and timely disclosure documents released by companies through the stock price exchanges on which they are listed, and treat them as stock price fluctuation factor information. These disclosures include content that may affect stock prices, such as financial results announcements, revisions of earnings forecasts, changes in business strategies, announcements of new products or services, and business alliances.

4. Target Stock Price Fluctuation Articles for Analysis

We focus on stock price fluctuation articles that contain both numerical stock price information and explanations of the price fluctuations based on corporate IR disclosures. Concretely, we utilize stock price data from the "ranking today" section provided by the stock information website Kabutan. This ranking targets stocks with the highest daily percentage increases and decreases, thereby enabling daily identification of stocks exhibiting significant price fluctuations.

In this paper, we target articles that include both stock price information and explanations of stock price fluctuations derived from the corresponding companies' IR disclosures. By aligning these articles with stock price fluctuation factor information (see the next section), we perform automatic generation of stock price fluctuation explanation articles.

5. Alignment between Stock Price Fluctuation Factors and Stock Price Fluctuation Articles

Given a stock price fluctuation article, for the designated company of the stock price fluctuation article, we collect corporate official disclosures concerning that company, then extract information that can be regarded as plausible factors explaining the observed stock price fluctuation. Specifically, we target corporate communications that were released prior to the publication date of the stock price fluctuation article, and establish correspondences based on semantic relatedness. In this alignment process, we allow not only exact lexical matches but also paraphrased or summarized expressions to be considered as corresponding information.

6. Article Generation from Stock Price Fluctuation Factors

In this study, we generate stock price fluctuation explanation articles using a large language model, taking as input stock price fluctuation factor information extracted from financial news and corporate disclosures, and recent numerical stock price data. This section describes the task formulation, input construction, prompt design, and generation strategy (Nishida and Utsuro, 2025).

The generation task in this study is formulated as a conditional text generation problem, where the model outputs a natural language stock price fluctuation explanation article conditioned on stock price fluctuation factor information and numerical stock price data. The input consists of the company name, numerical information including the direction and magnitude of the stock price fluctuation, and factor sentences that have been identified as causally related to the stock price fluctuation. The output is a short article written in a style comparable to that written by human financial journalists, concisely describing both the stock price fluctuation and its underlying causes. For numerical stock price information, we provide the closing prices over the most recent trading days and the magnitude of fluctuation, ensuring that the direction of the stock price fluctuation is clearly specified. For stock price fluctuation factor information, we use only those sentences extracted from financial news and corporate IR disclosures that are determined to have a causal relationship with the stock price fluctuation. This design prevents the inclusion of general industry descriptions or background information that lack direct causal relevance, thereby improving the precision of explanation generation.

In generating stock price fluctuation explanation articles, we employ GPT-4o as the LLM. The prompt explicitly instructs the model to act as a financial market journalist, to restrict the content to the provided input information, to avoid fabricating numerical values, to clearly state the direction of the stock price fluctuation, and to concisely describe the reasons for the change. In the zero-shot setting, only the task description and input information are

provided. In the few-shot setting, manually written example articles are included to guide stylistic and structural consistency. The temperature is set to 0 during generation to increase output determinism and ensure reproducibility in evaluation.

Under this framework, numerical data and textual factor information are integrated to generate explanations grounded in causal relationships. In particular, by explicitly providing stock price fluctuation factor information as input, the proposed method enables the generation of financial market articles with enhanced explainability and factual consistency.

7. Evaluation of Stock Price Fluctuation Articles

7.1. Overview

This section describes the human evaluation and the LLMs based automatic evaluation metrics employed to assess the stock price fluctuation articles generated by the LLM. In this study, ROUGE (Lin, 2004) is adopted as a baseline automatic evaluation metric. ROUGE is a lexical overlap based metric that measures similarity to reference texts; however, it is insufficient for evaluating the validity of explanations, such as whether the generated article appropriately captures the causes of stock price fluctuations. In particular, because stock price fluctuation articles typically contain a relatively small proportion of stock related terms within the entire text, ROUGE, which relies on lexical overlap, tends to produce unstable evaluations and is not necessarily well suited to this task. Therefore, this study proposes an evaluation metric based on LLMs, which has recently attracted considerable attention. We further analyze its correlation with human evaluation to demonstrate the effectiveness of automatic evaluation methods for stock price fluctuation article generation.

7.2. Manually Developing Reference Articles

As reference articles, 40 stock price fluctuation articles were manually written with reference to articles published on Kabutan between December 8 and 11, 2025. Each manually written stock price fluctuation article includes numerical information on stock prices, the primary cause of the price fluctuation, and supplementary explanations regarding the company or its products when necessary. The length of each article ranges from 150 to 350 Japanese characters, maintaining a level of conciseness comparable to that of actual stock price fluctuation articles. A concrete example is shown in Table 1. In describing the reasons for stock price fluctuations, we referred to the companies' disclosed IR information and based the descriptions on content judged to be directly related to the observed price fluctuations.

7.3. Evaluation Criteria

This section describes the evaluation criteria used to assess the quality of the generated stock price fluctuation articles. A stock price fluctuation article generally consists of multiple components, including (i) stock price information, (ii) an explanation of the factors underlying the price fluctuation, and (iii) supplementary information regarding the company or its business. Accordingly, we designed an evaluation scheme that separately assesses each of these components.

Specifically, human evaluation was conducted on a 10 point scale, divided into the following three categories:

- **Stock Price Information (4 points)**

  This criterion evaluates whether expressions related to stock price fluctuations, such as price increases or decreases, are correctly described and consistent with the actual price fluctuation.

- **Explanation of Stock Price Fluctuation Factors (3 points)**

  This criterion assesses whether the underlying factors behind the stock price fluctuation are accurately explained based on the company's disclosed information.

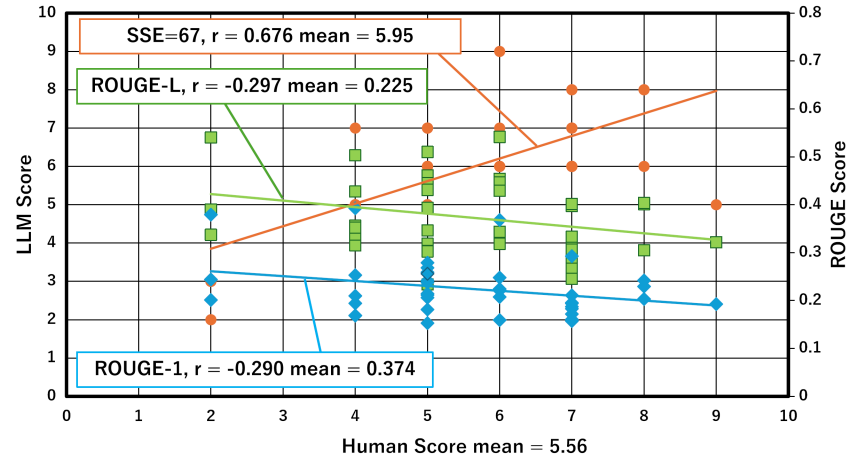
- **Additional Information (3 points)**

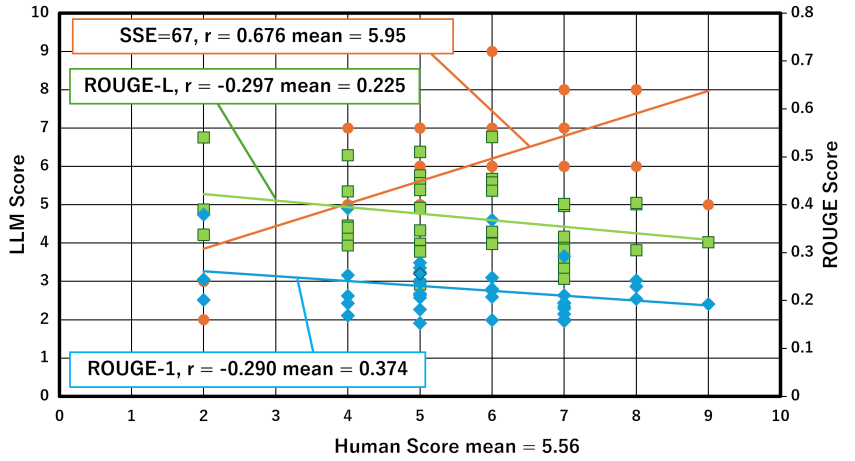
  Points are assigned based on either of the following types of information. The scores for these two types are not cumulative; instead, points are awarded only for the type that contains more substantial information: (i) a detailed explanation of the stock price fluctuation factors, or (ii) the accuracy of supplementary information, such as future outlooks or related business activities.

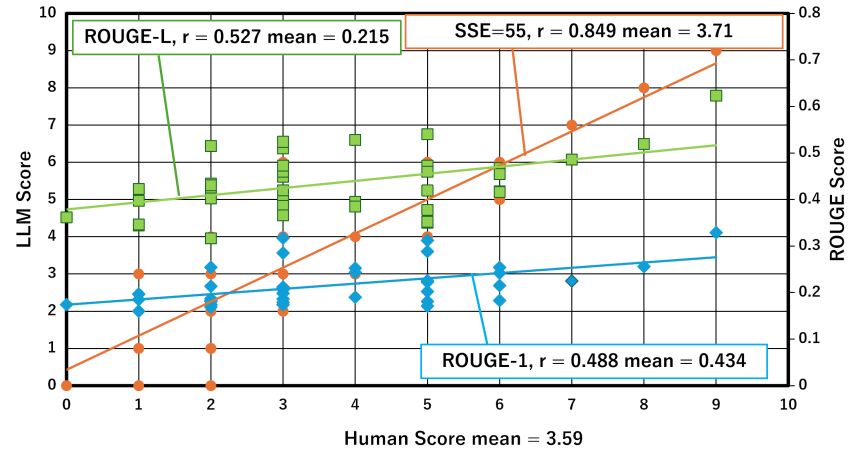
The total evaluation score for each article is calculated as the sum of the points assigned to each category. By separating the evaluation criteria in this manner, we aim to analyze not merely the textual similarity, but the extent to which the essential elements required in a stock price fluctuation article are satisfied.

7.4. Correlation Analysis between Human and Automatic Evaluation Results

In this section, we first generated 40 stock price fluctuation articles using GPT-4o (zero-shot and 5-







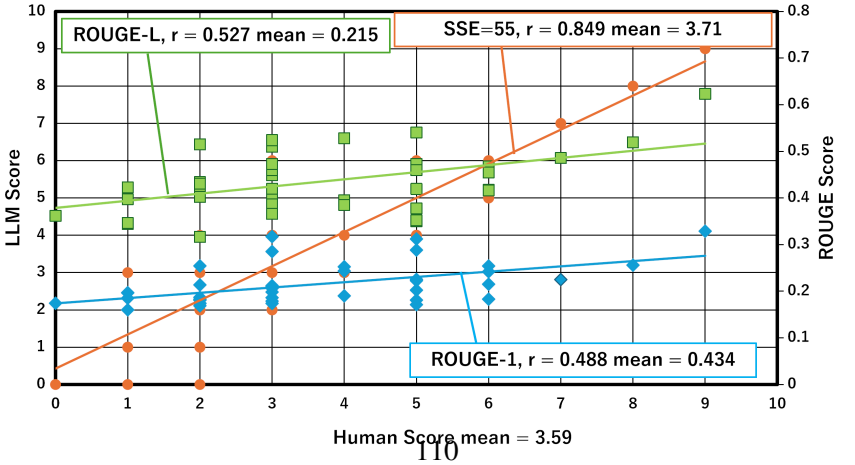


Figure 2: Correlation analysis results with human evaluation scores: Comparison of article evaluation results by LLM (Gemini 3, zero-shot) and ROUGE-1/L (F1 score).

Manually Written	Generated by GPT-4o (zero-shot)	Generated by GPT-4o (5-shot)
BrightPath Biotherapeutics <4594> continued its sharp rise. After the close of trading on the 8th, the company announced that it had received a notice of patent approval from the Japan Patent Office regarding a patent application related to its development pipeline BP1209 (fully personalized neoantigen vaccine). The patent concerns linker technology used in BP1209, a fully personalized cancer treatment neoantigen vaccine. This news was viewed as a positive catalyst.	The stock price of BrightPath <4594> rose from 50 yen to 52 yen and then to 55 yen over three days. The increase was attributed to the company's announcement that it had received a notice of patent approval from the Japan Patent Office regarding its development project BP1209. The patent relates to a fully personalized neoantigen vaccine for cancer treatment, and the technological progress was positively evaluated by the market.	The stock price of BrightPath Biotherapeutics <4594> is rising. The company announced that a patent related to BP1209 had received a notice of patent approval from the Japan Patent Office. This patent acquisition concerning BP1209, a fully personalized neoantigen vaccine for cancer treatment, is considered to have raised future expectations and supported the stock price.
Number of characters: 160	Number of characters: 163	Number of characters: 140
	Human: Stock price: 2 points Reason: 2 points Others: 1 point Total: 5 points	Human: Stock price: 2 points Reason: 2 points Others: 2 points Total: 6 points
	LLM: Stock price: 2 points Reason: 3 points Others: 1 point Total: 6 points	LLM: Stock price: 3 points Reason: 2 points Others: 1 point Total: 6 points
	ROUGE-1: 0.539, ROUGE-L: 0.458	ROUGE-1: 0.589, ROUGE-L: 0.521

Table 1: Example of a manually written article and articles generated by LLM (GPT-4o, zero-shot / 5-shot) (Article text, number of characters. Human and LLM (Gemini 3, zero-shot) evaluation results (Stock price: 4 points; Reason for fluctuation: 3 points; Others: 3 points; Total: 10 points), ROUGE-1/L (F1 score))

shot settings) and conducted a correlation analysis between human evaluation results and automatic evaluation results.

For automatic evaluation, we employed the evaluation criteria described in the previous section and used Gemini 3 provided by Google[4] as a zero-shot evaluator. By adopting the LLM-as-a-judge framework, the evaluator can comprehensively assess semantic consistency and logical coherence between the generated text and the reference text. Compared with ROUGE, which is based on n-gram overlap, this approach is more suitable for evaluating textual quality that involves structural and explanatory elements.

To measure the relationship between human and automatic evaluation results, we used the correlation coefficient r and the sum of squared errors (SSE). Let y_i and $\hat{y}_i$ $(i = 1, \ldots, n)$ denote the human evaluation score and the automatic evaluation score, respectively. SSE is defined as the sum of squared differences between the predicted values obtained from LLM based evaluation and the observed values obtained from human evaluation, as follows:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

A smaller SSE indicates a smaller discrepancy between the automatic (LLM based) evaluation and the human evaluation.

The correlation analysis results for the 40 reference articles are shown in Figure 2.

For the articles automatically generated by GPT-4o in the zero-shot setting, the LLM based automatic evaluation showed a strong positive correlation with human evaluation ($r = 0.676$), and the discrepancy from human scores was relatively small (SSE = 67). In contrast, ROUGE-1 ($r = -0.290$) and ROUGE-L ($r = -0.297$) exhibited negative correlations, indicating low consistency with human evaluation results.

For the articles generated by GPT-4o in the 5-shot setting, the correlation between automatic evaluation and human evaluation further improved ($r = 0.849$), demonstrating a very strong positive correlation. The discrepancy from human evaluation was also reduced compared to the zero-shot setting (SSE = 55). However, ROUGE-1 ($r = 0.527$) and ROUGE-L ($r = 0.488$) showed only moderate or lower levels of correlation.

These results indicate that, in the stock price fluctuation article generation task, the proposed method of using an LLM as an evaluator achieves a high correlation with human evaluation tendencies. In contrast, ROUGE was found to be inappropriate as an evaluation metric for stock price fluctuation articles.

[4] https://gemini.google/jp/about/?hl=ja

7.5. Case Study

This section presents concrete examples of stock price fluctuation articles automatically generated by GPT-4o in the zero-shot and 5-shot settings (Table 1) and analyzes the differences among human evaluation, LLM based evaluation, and ROUGE.

In the zero-shot setting, the generated article described specific numerical transitions in the stock price (50 yen $\rightarrow$ 52 yen $\rightarrow$ 55 yen), thereby supplementing numerical details. However, since the manually created reference article did not include explicit numerical transitions, this addition was judged as unnecessary information. As a result, the human evaluation assigned 2 points for stock price information, 2 points for the explanation of the price fluctuation factors, and 1 point for additional information, for a total of 5 points. Although the stock price expression was described merely as "increase," whereas the reference article used the term "continued rise," the direction of the fluctuation was consistent; therefore, 2 points were awarded. For the explanation of the price fluctuation factors, the reproduction of key terms such as "patent approval," "BP1209," and "fully personalized neoantigen vaccine" was positively evaluated. In comparison, the LLM based evaluation largely reproduced the human evaluation results, although it differed in assigning 3 points for stock price information.

In contrast, in the few-shot setting, the evaluation of stock price information and the explanation of fluctuation factors was the same as in the zero-shot case, but the appropriateness of supplementary information improved. Specifically, unnecessary numerical supplementation was suppressed, and the information was organized more coherently in accordance with the context. As a result, the human evaluation assigned 2 points for stock price information, 2 points for the explanation of fluctuation factors, and 2 points for additional information, for a total of 6 points. In the LLM based evaluation, differences were observed in two categories—3 points for stock price information and 1 point for additional information—yet the total score was consistent with the human evaluation.

8. Conclusion

In this paper, we proposed an automatic evaluation method for stock price fluctuation article generation using LLMs, based on the LLM-as-a-Judge framework (Chiang and Lee, 2023; Zheng et al., 2023), targeting the LLM based stock price fluctuation article generation method proposed in (Nishida and Utsuro, 2025). Specifically, the generated stock price fluctuation articles were evaluated automatically by an LLM under an item wise evaluation scheme.

In the evaluation experiments, we applied the proposed LLM-as-a-Judge-based automatic evaluation metric to stock price fluctuation articles automatically generated by GPT-4o in zero-shot and few-shot settings. The results demonstrated a sufficiently high correlation with human evaluation scores, indicating the effectiveness of the proposed automatic evaluation approach.

As future work, it will be necessary to establish an LLM based automatic evaluation method for stock price fluctuation articles generated from non-textual information sources, such as numerical data and chart data.

9. Bibliographical References

Sher Badshah and Hassan Sajjad. 2025. Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Proceedings of the 37th Advances in neural information processing systems*, pages 78142–78167.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Kentaro Imajo, Masanori Hirano, Shuji Suzuki, and Hiroaki Mikami. 2025. A judge-free LLM open-ended generation benchmark based on the distributional hypothesis. *arXiv preprint arXiv:2502.09316*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Shunsuke Nishida and Takehito Utsuro. 2025. Headline generation for stock price fluctuation articles. In *Proceedings of the 6th Workshop on Financial Technology and Natural Language*

Processing and Multi-Lingual ESG Impact Type Identification Shared Task, pages 184–195.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th Advances in neural information processing systems*, pages 46595–46623.

The Financial Document Causality Detection Shared Task (FinCausal 2026)

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, Sofía Roseti
Universidad Autónoma de Madrid
Laboratorio de Lingüística Informática
{antonio.msandoval, jordi.porta}@uam.es
{yanco.torterolo, maria.stanescu, melina.chatzi, sofia.roseti}@estudiante.uam.es

Abstract

The Financial Document Causality Detection shared task (FinCausal) is a competition organized within the Financial Narrative Processing (FNP) workshop series. It aims to identify the causal relationship between a question and its answer in a given financial context. The dataset is built from real annual reports drafted by Spanish IBEX 35 companies and several UK companies. The task includes two subtasks, one in English and one in Spanish. It is formulated as an Extractive Question-Answering (EQA) task in which, given a context (C) and a question (Q), participants must extract the verbatim answer span (A). The 2026 edition introduces several changes to increase task difficulty, including the reformulation of 10% of the questions to require deeper reasoning and a stronger emphasis on multi-step causal chains with three or more elements, achieved by removing overly simple cases and adding 500 new complex fragments per language. Another innovation is the adoption of an LLM-as-a-judge metric on a 1–5 scale, based on a rubric designed to align better with human preferences than Semantic Answer Similarity (SAS) and Exact Match (EM). This edition was hosted as part of the LREC conference in Palma de Mallorca, Spain.

Keywords: causal detection, EQA task, financial documents, LLM-as-a-judge

1. Introduction

The Financial Document Causality Detection Shared Task (FinCausal) is a long-running competition organized within the Financial Narrative Processing workshop series. The task focuses on text-internal causality in financial documents. Our objective is not to verify the factual truth of financial statements, but to evaluate how systems identify causes and effects as they are expressed in text. In its first editions, the shared task included only an English subtask (Mariko et al., 2021, 2022).

In the 2023 edition, causality detection was formulated as a span-extraction task (Moreno-Sandoval et al., 2023). Given a context and a span containing either a cause or an effect, participants had to extract the corresponding span that triggered it or was triggered by it. This edition also introduced the Spanish subtask. Evaluation relied on Exact Match (EM) at span level, and weighted F1, precision and recall at token level.

In 2025, the task shifted to an Extractive Question-Answering (EQA) task (Moreno-Sandoval et al., 2025). Given a context (C) and an abstractively formulated question (Q), participants had to extract the verbatim answer span (A) from the context. Because questions are abstractive while answers are extractive, this setup is more challenging for encoder-only systems and requires deeper contextual understanding. To better support generative models, Semantic Answer Similarity (SAS) (Risch et al., 2021) was introduced alongside EM to evaluate answers.

The 2026 edition builds on the 2025 framework and preserves the EQA formulation. This year, we prioritize *explanatory causality*, i.e., causes that lead to measurable effects, while reducing instances of *justificatory causality*, where text provides motives rather than direct triggers. In addition, EM and SAS were considered insufficient to capture all relevant response-quality nuances, and an LLM-as-a-judge framework was adopted. The main motivation is that many system outputs are semantically correct but lexically different from the reference, so a judge model can score semantic adequacy and causal grounding more faithfully than strict overlap-based metrics.

The 2026 design also increases task difficulty. We reviewed previous datasets and removed ambiguous or overly simple cases. We expanded the corpus with more than 500 new fragments per language, emphasizing multi-step causal chains with three or more elements. Moreover, 10% of the abstractive questions were rephrased to reduce lexical matching shortcuts and encourage deeper reasoning. Finally, training and test partitions were randomly re-split to distribute these changes evenly across the dataset.

2. Dataset

Building upon the 2025 dataset (Carbajo-Coronado et al., 2025), the current 2026 dataset (Moreno-Sandoval et al., 2026) transitions to a more complex EQA framework, distancing even more from the

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 114–124

Context	Question	Answer
<effect_2>Amadeus' non-air bookings declined by 1.5% in 2018 versus the previous year**</effect_2>** as a consequence of **<nested_cause_2><effect_1>**a decline in rail bookings**</effect_1>**, mostly driven by **<cause>**strikes impacting a key customer, which more than offset the double-digit increase in Amadeus' hotel bookings**</cause></nested_cause_2>**.	**<effect_2>**What factor caused Amadeus' 1.5% drop of non-air booking in 2018?**</effect_2>**	**<nested_cause_2>**a decline in rail bookings, mostly driven by strikes impacting a key customer, which more than offset the double-digit increase in Amadeus' hotel bookings**</nested_cause_2>**

Table 1: Sample for the English subtask marked with XML tags. Cause_1 corresponds to effect_1. They form nested_cause_2, corresponding to the effect_2. Effect_2 is used in the question (*Q*) to obtain the answer (*A*) from nested_cause_2.

2023 dataset (Moreno-Sandoval et al., 2023). The dataset is organized into two language-specific partitions, one for English and one for Spanish. Each example consists of a unique (*ID*), the context (*C*), the abstractive question (*Q*), and the gold-standard extractive answer (*A*). The distribution of the samples for the training and test partitions is detailed in Table 2.

The dataset for the Spanish subtask is sourced from the FinT-esp (Moreno Sandoval et al., 2020) corpus, which is composed of financial annual reports from Spanish IBEX 35 companies spanning 2014 to 2018. The official English translations of the 2018 reports were included and aligned, resulting in a bilingual ES-EN parallel corpus. For the English subtask, these English versions were combined with additional reports from the Lancaster UCREL research team corpus (El-Haj et al., 2019). All texts were manually annotated by linguists under expert supervision. The dataset is balanced across both subtasks to facilitate the development and evaluation of multilingual models. An example of the dataset is shown in Table 1. Additional information about the dataset and the competition is available on the official website.[1]

Subtask	Training Set	Test Set	Total
English	2,000	500	2,500
Spanish	2,000	503	2,503

Table 2: Distribution of samples for the FinCausal 2026 Shared Task.

3. Participants

A total of 20 teams registered for the task, 9 of which uploaded official submissions. All 9 participating teams submitted a corresponding system description paper for FinCausal. Among these teams, 7 of them competed in both the English and Spanish subtasks, while 2 of them focused exclusively on the English subtask. An additional team, while not submitting official test runs, provided a technical description of their proposed system, which is detailed in Subsection 5.4.

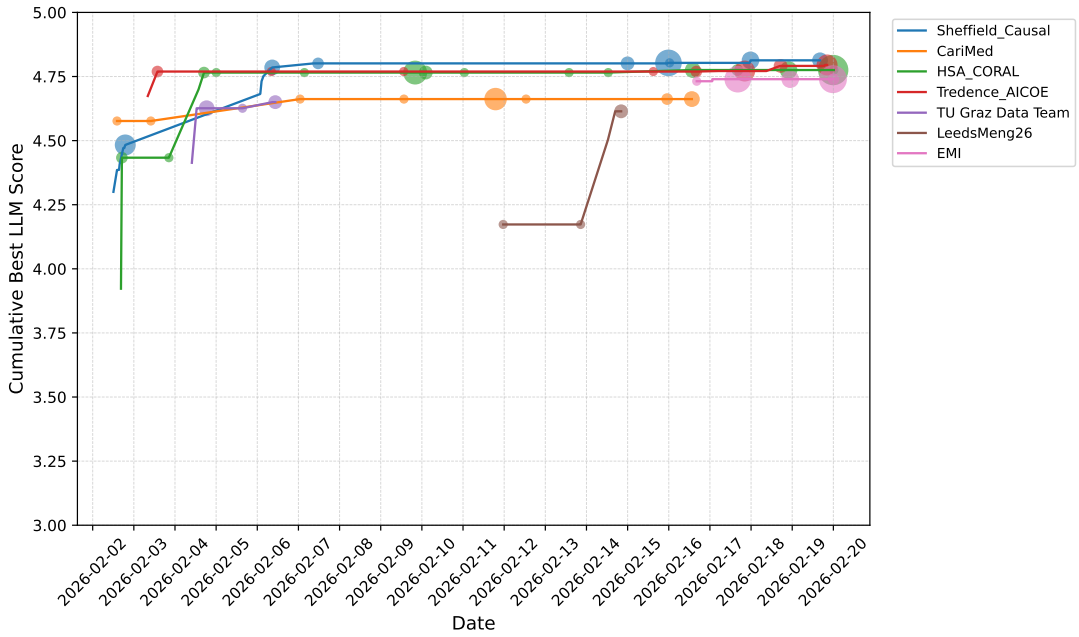
4. Competition Dynamics

In this competition, teams can make submissions that are scored using an internal metric not available to participants. Our goal in this section is to analyze the competition progress through the submission timeline and draw conclusions about competitive dynamics (e.g., participation pace, incremental improvements, stagnation periods, and convergence/divergence across teams). All data used for this analysis can be found in the competition leaderboard[2], which records each submission with the team name, score, and timestamp.

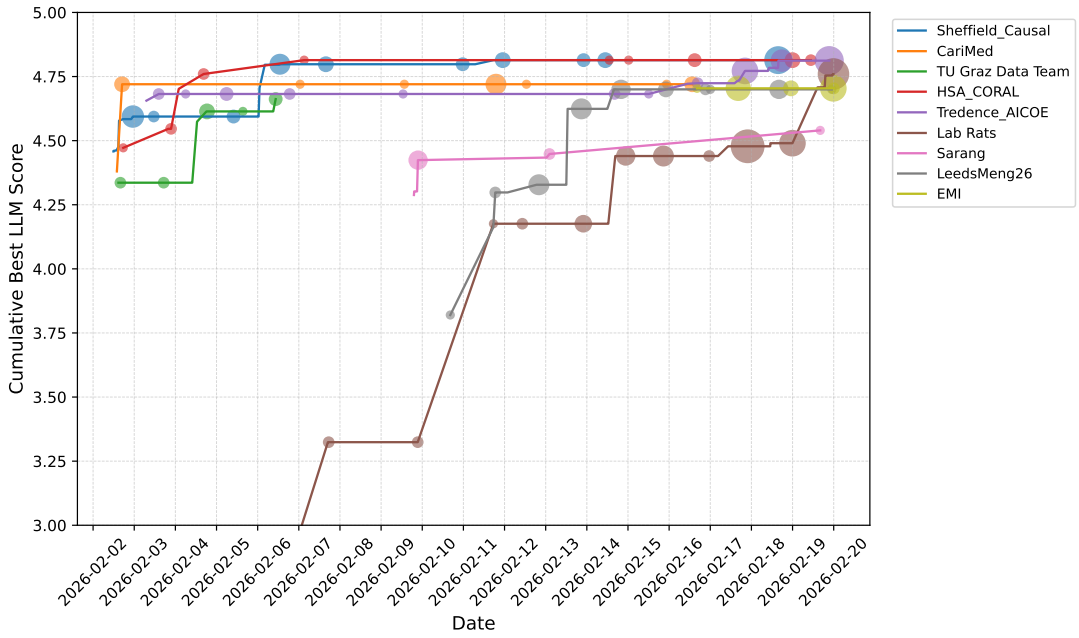
We build a time-ordered *timeline* that allows us to observe: the number of submissions per unit of time (activity), the evolution of each team's *best-so-far* (improvement curves), the evolution of the *leader frontier* (best overall score at each time), and the score distribution over time (convergence). Figure 1 shows two time-series line charts (Spanish on the left, English on the right) built from leaderboard records. Each colored line corresponds to one team and traces that team's cumulative best score over time, so upward steps indicate genuine improvements and flat segments indicate no improvement. Colors are used only to distinguish teams consistently across dates. The circles mark individual submissions at their submission timestamp and score. Circle size is proportional to submis-

[1] https://www.lllf.uam.es/wordpress/fincausal-26/

[2] https://leptis.lllf.uam.es/fincausal2026/



(a) Spanish Subtask



(b) English Subtask

Figure 1: Competition leaderboard dynamics for the Spanish (a) and English (b) subtasks. Each colored line represents one team and traces its cumulative best LLM score over time (step increases indicate improvements; flat segments indicate no improvement). Circles mark individual submissions at their timestamps and scores, and circle size is proportional to local submission density, so larger circles denote periods with more concentrated submission activity.

sion density at that point (i.e., larger circles indicate more submissions concentrated at that score/time region), helping visualize activity bursts and periods of low participation.

Temporal analysis reveals a consistent pattern across both subtasks: teams achieve rapid gains in the first days, followed by longer plateau phases with smaller marginal improvements. This behavior helps characterize iteration strategies (many submissions with small gains versus fewer submissions with larger jumps) and highlights *sprint* periods near the deadline. In the Spanish track, top systems

Team Name	Subtask	Core Methodology	Primary Model(s)
CariMed	Both	Multi-agent pipeline (RaR + RSA)	N/A
EMI	Both	Structural SFT (prompt repetition)	GPT-4.1-nano
HSA CORAL	Both	Prompt optimization, dynamic few-shot selection, and SFT	GPT-4.1-mini
Lab Rats	English	intra-context TF-IDF retrieval	Qwen3-4B-Instruct
LeedsMeng26	Both	Two-stage pipeline (candidate + verifier)	Qwen2.5 + Gemini
Sarang	English	Prompt optimization (DSPy/MIPROv2)	Gemma3-12B
Sheffield Causal	Both	Hybrid RAG (Dense) + bilingual SFT	GPT-4.1-mini
Tredence AICOE	Both	Voting-based ensamble (EN) | augmented SFT (ES)	GPT-5.1 / 4.1-mini
TU Graz Data Team	Both	SpanDiffusion (flow matching)	DeBERTa-v3
YT*	English	Difficulty-aware SFT + span anchoring	Llama-3.1-8B

Table 3: Comparison of participant methodologies. Teams are sorted alphabetically. *The YT team did not submit official results for the test evaluation phase.

converge early within a narrow score band (around 4.75–4.81), with limited late rank changes, suggesting faster stabilization. By contrast, the English track shows greater volatility, including later jumps from mid-ranked teams and stronger rank pressure near the top. Overall, the larger circles near the end of both timelines indicate denser submission activity, consistent with intensified last-minute experimentation.

5. Results

Table 4 reports, for each subtask, the best submission obtained by every participating team under the LLM-based evaluation protocol. For each entry, the table includes the final rank, team name, and best LLM score; it also reports summary statistics of score distribution, including the average score, to support direct comparison across language tracks. The Spanish and English rankings are presented side by side, making visible both top-level ties and score gaps across teams.

Overall, the results show strong and closely matched performance across teams in both subtasks. The English leaderboard is particularly competitive at the top, including a tie for first place. The Spanish leaderboard shows a similarly narrow gap among the leading systems. This distribution suggests that current systems are reaching a mature performance regime in this competition.

5.1. Spanish Subtask

The results for the Spanish subtask, as detailed in Table 4, reflect a high degree of technical proficiency. While the top three contenders remained consistent with the English track, the internal hierarchy shifted slightly. Sheffield Causal emerged as the winner with a score of 4.813, closely followed by Tredence AICOE and HSA CORAL. The overall performance was notably homogeneous; the gap between the leading submission and the bottom of the leaderboard did not exceed 0.2 points, with all participants scoring above the 4.6 threshold. This narrow variance suggests that the complexity of the Spanish financial corpus was effectively addressed by the participants' multilingual strategies.

5.2. English Subtask

The performance across the English subtask reveals an exceptionally competitive landscape. A remarkable tie for the first position was achieved by HSA CORAL and Sheffield Causal, both reaching a near-perfect score of 4.814 out of 5. The margin for the top positions was minimal; Tredence AICOE secured the third spot, trailing by a mere 0.002 points. This tight clustering of results extends throughout the ranking, where even the lowest-scoring submissions maintained a high standard of causal extraction, demonstrating the robustness of current LLM-based architectures in processing English financial narratives.

5.3. System Descriptions

To contextualize the ranking results, we briefly summarize the system-design choices reported by participants. Table 3 provides an overview of the systems employed by each team.

5.3.1. English and Spanish Systems

Team **Carimed** (Jay et al., 2026) introduced the VERSA system to tackle the task. It is a five-stage multi-agent pipeline in a zero-shot scenario that relies on API-connected models. It utilizes

Rank	Team Name	LLM Score					Average
		1	2	3	4	5	
1	Sheffield Causal	1	2	16	52	432	4.813
2	Tredence AICOE	0	2	28	41	432	4.795
3	HSA CORAL	4	1	21	52	425	4.775
4	EMI	2	6	28	49	418	4.739
5	CariMed	14	7	18	57	407	4.662
6	TU Graz Data Team	3	7	25	93	375	4.650
7	LeedsMeng26	8	7	24	83	379	4.614
	Total	**32**	**32**	**160**	**427**	**2868**	

(a) Spanish Subtask

Rank	Team Name	LLM Score					Average
		1	2	3	4	5	
1	HSA CORAL	3	2	21	33	441	4.814
1	Sheffield Causal	4	2	18	35	441	4.814
3	Tredence AICOE	1	3	24	33	439	4.812
4	Lab Rats	3	9	24	33	431	4.760
5	CariMed	5	6	28	46	415	4.720
6	EMI	6	6	40	26	422	4.704
7	LeedsMeng26	3	11	26	53	407	4.700
8	TU Graz Data Team	8	9	58	55	370	4.662
9	Sarang	1	3	24	33	439	4.540
	Total	**34**	**51**	**263**	**347**	**3805**	

(b) English Subtask

Table 4: LLM-based ranking of each team's best submission for the Spanish (a) and English (b) subtasks. Each row reports the team rank, team name, score distribution and the LLM score (avg.) of their best submitted system.

two main techniques: (1) Rephrase-and-Respond (RaR), where the model optimizes the prompt; and (2) Recursive Self-aggregation (RSA), which provides candidate answers and recursively adds subsets to obtain consensus. Stage 1 functions as a causal analyst, stage 2 applies the rephaser (RaR), stage 3 extracts the relevant information, stage 4 aggregates candidates (RSA), and stage 5 validates the candidates before providing the final answer. This approach obtained a score of 4.6620 in Spanish (5th) and 4.720 in English (5th).

Team **HSA CORAL** (Gautam et al., 2026) compares three approaches to address the EQA: (1) encoder-only for token classification using BERT-based models, (2) encoder-decoder for sequence-to-sequence generation using BART-like models, and (3) decoder-only models for generation, enforcing extraction with prompt optimization, few-shot selection, and fine-tuning. They retrieve the most similar cosine-similarity-based C and Q pairs from the training set to those pairs from the test set to include them as shots. They found that fine-tuned generative models outperform encoder-only and encoder-decoder models. Also, 20-shot prompting enables compact models to outperform bigger models in a zero-shot scenario. Their best system consists of a bilingually fine-tuned GPT-4.1 mini with 20 shots similar to the test sample, achieving a score of 4.7753 in Spanish (3rd) and 4.814 in English (1st).

Team **Tredence AICOE** (Chopra et al., 2026) presented a multilingual financial causality extraction system for English and Spanish based on few-shot prompting, supervised fine-tuning, data augmentation, and ensemble arbitration. They experimented with models such as GPT-5, Gemini 3.0 Pro, Qwen-14B, Gemma-12B, GPT-4.1 mini, GPT-OSS 20B, and GPT-5.1, and explored techniques including joint EN+ES training, LoRA-style efficient fine-tuning, bidirectional translation, synthetic data generation, distilled chain-of-thought reasoning, confidence-based selection, semantic consensus, and voting-based ensembles. Overall, the best results come from multilingual fine-tuning plus selective ensembling, with GPT-5.1-arbitrated voting performing best in English and synthetically augmented GPT-4.1 mini performing best in Spanish. In Spanish, the best score was 4.795 (2nd), achieved by GPT-4.1 mini fine-tuned with synthetic data augmentation. In English, the best result was

obtained with a voting-based ensemble arbitrated by GPT-5.1, which achieved 4.812 (3rd), outperforming all standalone prompting and fine-tuned systems.

Team **EMI** (Attak, 2026) presented a system based on supervised fine-tuning of instruction-following language models, with particular emphasis on prompt repetition as a training strategy to reinforce the relationship between the question format and the expected extractive answer. The authors evaluate both open-weight (Qwen2.5-7B, Qwen2.5-14B) and proprietary models (GPT-4.1-Nano), and show that this simple intervention can improve extraction fidelity, especially for open models, by reducing over-generation and helping the model stay closer to the source span. Their best systems (GPT-4.1-nano) obtained 4.7396 in Spanish (4th) and 4.704 in English (6th).

Team **LeedsMeng26** (Shahrouri et al., 2026) presented a two-stage extractive question answering system for FinCausal 2026. They cast financial causality detection as a QA problem over English and Spanish financial texts, returning a verbatim span from the context rather than generate a free-form answer. Their system works in two steps. First, a model generates an initial candidate span under strict prompting designed to force extractive behavior. Then a second model acts as a verifier and boundary refiner, checking whether the candidate is correct and adjusting its span boundaries when necessary, while still being constrained to output only a contiguous substring from the source text. The paper says this second stage was introduced to fix typical span errors such as truncation, overrun, or occasional paraphrasing. Their final submitted configuration was Qwen-2.5-1.5B-Instruct + Gemini 2.5-flash achieveving scores of 4.6143 in Spanish (7th) and 4.7000 in English (7th).

Team **TU Graz Data Team** (Niess and Kern, 2026) introduced SpanDiffusion to approach financial causal question answering as an extractive span prediction problem, but replaced standard start–end classification with a continuous denoising approach in a system called SpanDiffusion. The question and context are encoded with DeBERTa-v3-large plus LoRA, and the target answer is represented as two Gaussian masks marking the start and end of the span. A transformer denoiser trained with flow matching reconstructs these masks from noise, and the final span is obtained by taking the peak of each mask. This design aims to model boundary uncertainty while preserving the extractive nature of the task. In the results, the proposed method proves competitive but not superior to a simpler baseline, namely DeBERTa-v3-large + LoRA + a linear span head. The best SpanDiffusion model reaches 83.0 Exact Match, with notable gains from LoRA and from using flow matching instead of DDPM. However, the standard span-classification baseline achieves 85.8 Exact Match, outperforming the diffusion model with much lower complexity. Overall, the paper's contribution is therefore mainly methodological, offering an original alternative to conventional extractive QA rather than a stronger empirical system. Their best systems achieved a score of 4.6501 in Spanish (6th) and 4.662 in English (8th).

Team **Sheffield Causal** (Alqarni et al., 2026) addressed financial causal question answering in English and Spanish as a generative extractive task based on GPT-4.1-mini. The authors combine prompt engineering, retrieval-augmented generation (RAG), and supervised fine-tuning, while enforcing strict verbatim extraction from the input context. Their pipeline has three stages: indexing the training set, retrieving top-k similar examples, and constructing few-shot prompts for either the base or the fine-tuned model. They compare several retrieval strategies, including random example selection, BM25, dense retrieval with text-embedding-3-large and LlamaIndex, a pattern-aware retrieval method that groups questions into CAUSE, EFFECT, or OTHER templates, and a hybrid BM25+dense approach using reciprocal rank fusion. In addition, they evaluate simple, expert, and multilingual prompts, and fine-tune GPT-4.1-mini on bilingual training data formatted as question-context-answer triples. Their system ranked first in both subtasks, reaching a score of 4.813 for Spanish (1st) and 4.814 for English (1st).

5.3.2. English-only Systems

Team **Lab Rats** (Sarda et al., 2026) participation in the competition consists of two main aspects: QLoRA SFT of Qwen3-4B-Instruct on the English dataset, which was adapted to the ChatML instruction format required by the model; and an intra-context TF-IDF retrieval to enrich context for enforcing verbatim span extraction at inference time. The inference pipeline involves four stages: (1) loading the model in 4-bit, (2) intra-context retrieval, comparing each sentence from the given C against the Q, thus filtering the most relevant fragment, (3) constrained prompting, where the model is instructed to extract the span verbatim, and the previously retrieved fragment is also provided, and (4) greedy decoding and deterministic settings aimed at improving extraction accuracy. They achieved a score of 4.760 in English (4th).

Team **Sarang** (Trivedi and Chindukuri, 2026) presented a system based on few-shot prompting and automated prompt optimization for Gemma3-12B. Using DSPy and the MIPROv2 teleprompter, the system optimizes instructions and demonstration examples drawn from the training set, and performs inference locally through Ollama. The paper

compares this setup with RoBERTa and DeBERTa baselines finetuned with other QA datasets and with other few-shot LLM configurations, finding that Gemma3-12B with medium prompt optimization performs best. Their top system obtained an LLM Score of 4.540 in the English shared sub-task (9th), highlighting the effectiveness of prompt optimization for financial causal QA.

5.4. Additional English-only System

Team **YT** utilized a LoRA-finetuned Llama-3.1-8B model using a difficulty-aware training strategy. Their methodology involves a custom labeling scheme that categorizes instances into simple, chain, and abstractive types based on structural and semantic complexity. To ensure strict verbatim extraction, the system employs a three-level span-anchoring mechanism—combining case-insensitive matching, semantic sentence retrieval, and re-prompting—complemented by targeted post-processing rules to truncate redundant clauses. Although the team did not submit official runs for the shared task, their internal evaluations on a partitioned subset of the 2026 dataset demonstrated the effectiveness of combining difficulty-stratified training with post-processing.

6. Evaluation

In FinCausal 2023 (Moreno-Sandoval et al., 2023), the task focused on identifying cause–effect relations linked to events or quantified facts in financial texts. Because it was formulated as an extractive task, system performance was assessed with EM to quantify the proportion of predictions that exactly match the reference span, and token-level precision, token-level recall, and weighted F1 to quantify partial overlap quality for extracted cause and effect spans.

In FinCausal 2025 (Moreno-Sandoval et al., 2025), the extraction task was reformulated as a question-answering task in which questions about causes or effects are posed, and system responses are evaluated with EM and semantic answer similarity metrics. This change accommodates the growing use of generative prompting-based models, many of them based on GPT architectures. For these models, a strict lexical metric such as EM alone is often insufficient, because generative systems can produce answers that are semantically correct but phrased differently from the references. SAS measures semantic similarity between texts rather than exact lexical overlap, making it well suited for abstractive generation tasks. The metric represents texts as vector embeddings using pre-trained models such as BERT (Devlin et al., 2019) or Sentence Transformers (Reimers and Gurevych, 2019), and computes cosine similarity between them. This allows the evaluation to capture cases where two answers express the same meaning despite differences in wording or structure.

In the FinCausal 2026 edition, system submissions are evaluated through an LLM-as-a-judge framework. For each system prediction, the judge model receives a fixed evaluation rubric and scores the response according to a uniform set of criteria. Specifically, each answer is rated on a five-point Likert scale, whose levels capture different degrees of semantic alignment between the predicted answer and the gold-standard reference. The full FinCausal 2026 rubric is provided in Appendix A. In our setup, the judge model is `openai/gpt-oss-20b` (OpenAI, 2025), a 20-billion-parameter open-weight language model from OpenAI's GPT-OSS family released in August 2025. We use the model with a medium reasoning configuration and explicitly instruct it in the prompt to generate concise score justifications, thereby improving the transparency and auditability of the evaluation procedure.

6.1. Error Analysis

The following initial analysis is based on a linguistic review of model outputs graded from 1 to 5. The predictions from the best system of each participant were observed, thus enabling the identification of common error patterns found in the lower and middle scoring ranges. Superficially, no errors were found on the scores of 5, but a more thorough analysis should be performed.

Score 1: Structural failure. At this level, models show major structural problems. While they often identify causal markers (like "due to" or "thanks to"), they extract the wrong information or focus on unrelated events nearby. Another common issue is empty referencing, where the model only extracts the connector (e.g., "Due to the above") without the actual explanation. Additionally, models at this stage sometimes reverse the cause and effect or skip essential steps in a causal chain. These models are also prone to neglecting or hallucinating quantitative financial data; a failure severely penalized by the LLM judge, which assigns the minimum score to responses lacking numerical precision.

Score 2: Incomplete and inaccurate. Errors here involve missing information or technical mistakes. Models often cut the answer too short, leaving out the specific details that provide the actual argument. They also tend to confuse similar financial terms or acronyms (like DVA instead of CVA). We also observed a metric bias: LLM judges sometimes give higher scores to fluent-sounding answers, even if the content is partially incorrect.

This behaviour is typical of metrics based in neural models (Kovacs et al., 2024; Freitag et al., 2024).

Score 3: Partial and selective. These responses overlap with the correct answer but are incomplete. Models often pick only the first factor in a list and ignore the others, providing a one-sided view of the event. They also frequently leave out important time-related details (like specific dates or years). While the core reason is usually captured, these models tend to ignore secondary details or consequences, resulting in an over-simplified version of the reference.

Score 4: Precise with minor noise. A score of 4 represents an accurate answer that includes unnecessary noise. The most common error is marker dragging, where the model includes extra words like connectors or introductory verbs (e.g., "resulting in" or "allowing") that were not part of the target answer. In other cases, the model adds extra context that is true but goes beyond the exact boundaries set by the experts.

The errors progress from total confusion at score 1 to minor noise at score 4. The main challenges for these models are navigating complex financial texts with many variables and staying within the strict boundaries of the required answer. Finally, the metric bias seen in scores 2 and 3 suggests that LLM judges often favor smooth, fluent writing over literal, word-for-word accuracy. Therefore, the differences between scores of 2 and 3 are often blurry, while the scores of 1 and 4 seem reliable.

7. Conclusions

This FinCausal 2026 edition had higher participation than previous editions in terms of effective system-description submissions. Consequently, it enabled us to gain a clearer overview of the current state of causal EQA in the financial domain. Some insights can be drawn from the results obtained by participants.

Bilingual fine-tuning performs consistently better than monolingual fine-tuning, as shown by direct comparisons in the HSA CORAL and Tredence AICOE submissions. Additionally, as highlighted by HSA CORAL, smaller fine-tuned models can outperform larger zero-shot models.

Another clear trend was the use of retrieval-based techniques and semantic matching. Four teams used similarity measures at different stages of their pipelines for different purposes. HSA CORAL applied vector similarity at inference time to retrieve few-shot examples similar to the test C and Q for in-context learning. Sheffield Causal followed a similar approach to HSA CORAL but conducted a broader evaluation of retrieval methods, including hybrid BM25 and dense retrieval. In contrast, Lab Rats performed intra-context filtering by comparing each sentence in the provided C with the Q to isolate the most relevant fragment before generation. Finally, YT integrated semantic similarity at two points: for the initial difficulty-based classification of the dataset and for a span-anchoring mechanism during post-processing.

On a different note, complex pipelines that rely on zero-shot settings still appear insufficient to consistently enforce verbatim extractive answers, as seen in CariMed's VERSA system and Sarang's prompt-optimization strategy. Compared with similarly complex pipelines that include fine-tuning, such as Tredence AICOE's, these results suggest that fine-tuning remains essential for stronger performance.

Regarding the main model choice, GPT-4.1 mini was used by several teams and proved to be a strong option despite its size. Sheffield Causal used it to obtain first place in both subtasks, tying with HSA CORAL in English. HSA CORAL also used GPT-4.1 mini in its best-performing system, and Tredence AICOE used it for their best Spanish run. A smaller variant, GPT-4.1 nano, was used by EMI. These results suggest an advantage for proprietary models in this edition. Given the compact size of the mini and nano variants, they also offer an attractive cost-performance ratio. On the open-source side, Qwen2.5 appeared in the systems of LeedsMeng26, while Qwen3 appeared in Lab Rats'. Gemma 3 was used by Sarang, while TU Graz Data Team used DeBERTa-v3 together with a span diffusion approach; YT used Llama-3.1-8B.

To conclude, despite the difficulty of the task, all participants achieved strong scores. This highlights both the quality of the participants' work and the consistency and coherence of the dataset in both languages. In addition, the shift to an LLM-as-a-judge metric has proved useful in light of the error analysis. However, a deeper future analysis of its decisions and of its alignment with human judgment is still needed. For future FinCausal editions, some important changes should be considered to maintain participant interest; for instance, including an additional subtask formulated as a pure QA task with abstractive answers, while preserving the current EQA approach.

Acknowledgments

We would like to thank FinCausal 2026 participants for their outstanding contributions to this shared taks.

This work is framed under the Spanish National Project GRESEL (PID2023-151280OB-C21). It was also partially funded by grant PTA2023-023812-I (awarded to Yanco Amor Torterolo Orta) through MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+).

8. Bibliographical References

Aali Abdullah Alqarni, Mark Stevenson, and Arif Dwi Laksito. 2026. A Comparative Study of RAG Approaches and Fine-Tuning for Causal QA in Financial Text. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Sanae Attak. 2026. Improving Verbatim Financial Causality Extraction with Supervised Fine-Tuning and Prompt Repetition. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Ankush Chopra, Shubham Sharma, and Ashmani Kumar. 2026. Extracting Financial Causality: A Multilingual Approach with SLM Fine-Tuning and LLM-Arbitrated Ensembles. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Akash Kumar Gautam, Serhii Hamotskyi, and Christian Hänig. 2026. Causal Connections: Leveraging Multilingual Fine-Tuning for Financial QA@FinCausal 2026. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Aldan Jay, Rafael Berlanda, Yoelvis Moreno, and Vincent Santamarta. 2026. VERSA: Verbatim Extraction via Rephrasing and Self-Aggregation for Financial Causality. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The Financial Document Causality Detection Shared Task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The Financial Causality Extraction Shared Task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, Jordi Porta-Zamorano, Yanco-Amor Torterolo-Orta, and Doaa Samy. 2025. The Financial Document Causality Detection Shared Task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno Sandoval, Ana Gisbert, and Elena Montoro. 2020. FinT-esp: A corpus of financial reports in Spanish. In Miguel Fuster-Márquez, Carmen Gregori-Signes, and José Santaemilia Ruiz, editors, *Multiperspectives in analysis and corpus design*, pages 89–102. Comares, Granada.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The Financial Document Causality Detection Shared Task (FinCausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Georg Niess and Roman Kern. 2026. SpanDiffusion: Flow Matching over Continuous Span

Masks for Financial Causal Question Answering. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic Answer Similarity for Evaluating Question Answering Models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bavya Sarda, Pulkit Chatwal, and Sonal Dabral. 2026. QRAFT: QLoRA Retrieval-Augmented Fine-Tuning for Causal Span Extraction in Financial Documents. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Zaid Shahrouri, Ayomide Ivienagbor, Idrees Asad, Rijul Shrestha, Yasemin Bal, and Zahaab Nadeem. 2026. LeedsMEng26: Qwen + Gemini for FinCausal 2026 Causality Detection in Financial Narrative Texts. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Avinash Trivedi and Mallikarjuna Chindukuri. 2026. Financial Causal QA via Instruction and Prompt Tuning of Gemma3-12B. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

A. FinCausal 2026 Rubric

```
You are an expert evaluator. Your task is to rate how good an STUDENT_ANSWER is
for a given QUESTION, a CONTEXT and a REFERENCE_ANSWER according to
the rubric below.

RUBRIC (score from 1 to 5):
- 5: Excellent quality: The prediction is semantically identical or fully equivalent to the gold standard response. Minor formal variations (e.g., punctuation, casing) are tolerated. The content perfectly addresses the causal question, showing no signs of omission or irrelevant inclusion. The answer is deemed fully appropriate and reliable.
- 4: Good quality: The prediction matches the gold-standard answer in full but included small additional content from the context. These additions did not compromise the semantic correctness of the answer but extended it slightly. Therefore, the prediction could be seen to be complete, relevant and informative, although wordy.
- 3: Medium quality: The prediction contains the central idea or a correct causal link, but is either incomplete or diluted with unrelated information. It may capture the start of a causal phrase but miss important qualifiers or follow-up clauses. Alternatively, the prediction might include correct content but extend unnecessarily beyond the relevant span.
- 2: Low quality: The predictions demonstrates only a superficial connection to the question or to the correct answer. Typically, large portions of the expected content are omitted, and irrelevant elements may have been added. The response might contain a partial clue or topic-related phrase, but failed to provide a clear, informative, or accurate answer. This category captures both underinformative and noisy outputs.
- 1: Very poor quality: Predictions in this category failed entirely to answer the question. These responses were irrelevant, incorrect, or confusing, and often exhibited no meaningful overlap with the gold-standard answer. Even if some surface text matched the source, the essential causal content was absent.

First, briefly explain your reasoning.
Then assign a single integer score from 1 to 5.

Return your response as pure JSON with this exact schema:

{{
```

```
  "score": <integer 1-5>,
  "reasoning": "<short explanation>"
}}

CONTEXT:
{context}

QUESTION:
{question}

REFERENCE_ANSWER:
{reference}

STUDENT_ANSWER:
{answer}
```

B. Language Resource References

Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, and Paula Gozalo. 2025. The Financial Document Causality Detection Shared Task (FinCausal 2025): Dataset.

Mahmoud El-Haj, Steven Young, and Paul Rayson. 2019. Annual reports key sections corpora 2003 to 2017.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, and Jordi Porta. 2023. The financial document causality detection shared task (FinCausal 2023): Dataset.

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

Sheffield NLP at FinCausal 2026: A Comparative Study of RAG Approaches and Fine-Tuning for Causal Q&A in Financial Texts

Aali Alqarni, Mark Stevenson and Arif Laksito
School of Computer Science, University of Sheffield
Sheffield, United Kingdom
{aaalqarni1, mark.stevenson, alaksito1}@sheffield.ac.uk

Abstract

This paper describes our approach to the FinCausal 2026 shared task, which addresses causal question answering from financial documents in English and Spanish. We investigated the effectiveness of fine-tuned generative models combined with Retrieval-Augmented Generation (RAG). Our approach compares five retrieval strategies across base and fine-tuned GPT models (GPT-4.1-mini). RAG-based few-shot selection performed better than random sampling, particularly for the base model. In the FinCausal 2026 official run, this approach was ranked first in both the English and Spanish sub-tasks, obtaining LLM scores of 4.8140 and 4.8131 out of 5, respectively.

Keywords: Question and Answering (Q&A), Causality, Large Language Model (LLM), Generative Pre-trained Transformer (GPT), Retrieval-Augmented Generation (RAG)

1. Introduction

Financial documents, including earnings reports and financial analysis articles, contain rich causal relationships that drive market movements and trends. Understanding causality in these documents is fundamental for risk assessment, investment decision making, and market analysis (Cormack et al., 2009). Manual extraction of these relations is time-consuming and labour intensive. Automated methods offer a solution but face challenges due to the complexity of financial language and the diversity of causal expressions (Li et al., 2024; Cao et al., 2022).

To advance research in this area, the FinCausal 2026 shared task (Moreno-Sandoval et al., 2026) builds on the 2025 edition, which introduced a question and answering (Q&A) framework for extracting causal relationships from financial texts (Moreno Sandoval et al., 2025). The core objective of the 2026 edition remains the identification of events that explain financial outcomes such as revenue changes, market movements, or corporate decisions, while also introducing more rigorous benchmarks. The dataset has been expanded with over 500 complex causal cases, including multi-element causal chains, and questions have been rephrased to reduce reliance on sentence-level similarity. The task has also shifted to LLM-as-a-judge evaluation rather than the Exact Match (EM) and Semantic Answer Similarity (SAS) measures used previously (Risch et al., 2021).

1.1. Task Description

Objective: Given a natural language question, Q, and corresponding financial text context, C, systems must extract a verbatim span, A, from C that captures the underlying causal relationship.

Example:

Q: "What explains their strong performance in health and safety?"

C: "As a result of **these very high standards and relentless focus**, we have a strong performance in health and safety" (A indicated by **bold font**).

2. Related Work

The FinCausal shared tasks have been organised as a benchmark for evaluation methods that detect causal relationships in financial texts. Early editions (2020-22) formulated causality detection as either causal sentence classification or cause–effect span identification, with most systems relying on extractive models such as BERT and BiLSTM-CRFs (Mariko et al., 2020, 2022). The 2023 edition extended the task to a multilingual setting by requiring systems to identify causal relationships in financial texts written in English and Spanish. This edition marked a notable transition to the use of generative large language models (LLMs) such as GPT (Moreno-Sandoval et al., 2023). The 2025 edition reframed the task as a Q&A problem, requiring systems to extract answer spans from financial contexts given a natural language question.

LLM-based causal extraction. Recent work in cause–effect span detection tasks has explored GPT-based approaches for causal extraction from financial texts. Shukla et al. (2023) combined Retrieval-Augmented Generation (RAG) with few-shot prompting using GPT-4, while the LTRC IIITH team explored chain-of-thought (CoT) prompting to enhance reasoning performance (Moreno-Sandoval et al., 2023). These approaches suggested that in-context learning can effectively iden-

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 125–131

ID	Context	Question
English		
29	**At the start of the year, we also reduced the size of our transport fleet by 10% in light of network changes. At this point, we decided against further reducing the size of the fleet due to the future demand from new contracts**. As a result, the business carried significant excess costs of under-utilised trucks during the current financial year.	What accounts for the business carrying significant excess costs derived from under-utilized trucks during the current financial year?
Spanish		
76	Por el contrario, **la Comunidad de Estados Independientes registró un descenso de la producción del 28%. Esta reducción fue menos acusada que la de la demanda interna** debido a las exportaciones.	¿Qué efecto tuvieron las exportaciones?

Table 1: Example entries from the English and Spanish datasets. Answers are highlighted in **bold** within the context.

tify cause–effect relationships without fine-tuning.

LLM-based Q&A. Niess et al. (2025) compared the performance of extractive models, such as BERT, with generative LLMs, including Llama, in zero-shot and few-shot settings for causal Q&A. Their findings indicated that instruction-tuned LLMs without task-specific fine-tuning were competitive. However, a fine-tuned Llama model trained on a multilingual dataset significantly outperformed both approaches.

Beyond stand-alone model comparisons, recent work has explored the integration of LLMs within retrieval-augmented Q&A frameworks to enhance factual grounding and reduce hallucinations. For instance, Yang et al. (2026) proposed Structured-Semantic RAG (SSRAG), a hybrid architecture that incorporates LLM-based query augmentation, prompt-guided source routing, and unified graph–vector retrieval to improve answer faithfulness across open-domain Q&A benchmarks. In the medical domain, Aljohani and Alsanoosy (2026) proposed a modular hybrid RAG framework that integrates sparse retrieval (BM25) with dense biomedical retrieval (MedCPT; Jin et al. 2023) to enhance medical Q&A. Their approach demonstrated substantial improvements in retrieval recall, precision, and generation faithfulness across PubMedQA, MedMCQA, and MedQA-US benchmarks (Pal et al., 2022; Jin et al., 2019, 2021). Despite these advances, limited work has systematically examined hybrid retrieval approaches for domain-specific causal Q&A in financial texts, especially in multilingual settings. This gap highlights the need for retrieval strategies that are specifically designed for causal financial Q&A tasks.

3. FinCausal 2026 Dataset

The FinCausal 2026 dataset (Moreno-Sandoval et al., 2026) contains English text from UK financial reports dated 2017 and Spanish text from Spanish financial reports dated 2014 to 2018. Entries consist of an abstractive question asking about a cause or effect, a context passage, and a verbatim answer extracted from the context.

The dataset includes different types of causality. Some have explicit causal markers such as 'due to' and 'because', while others require reasoning from the context to identify implicit causal connections. It includes complex cause-and-effect relationships, such as causal chains of three or more elements where multiple events are linked sequentially (e.g., A causes B, which leads to C). Questions can be classified into cause-seeking (e.g., 'What caused the...?'), effect-seeking (e.g., 'What was the impact of...?'), and others that do not follow a specific causal pattern (e.g., 'What did these considerations entail?').

Dataset Description. Context passages vary in length from a single sentence to multiple sentences, and answer spans range from a short phrase to multiple clauses. Table 1 shows representative examples from the English and Spanish datasets.

Dataset Statistics. The dataset comprises 2,000 labelled training instances per language, with 500 blind test instances for English and 503 for Spanish. We applied an 80:20 split to the training data. This resulted in 1,600 instances per language for fine-tuning and retrieval indexing, and 400 per language for development (including ablation studies and prompt engineering).

4. Methodology

4.1. Model Selection

A generative approach was adopted following previous work which demonstrated that fine-tuned generative models notably outperform traditional extractive Q&A systems (Moreno Sandoval et al., 2025). Figure 1 shows the overall system architecture.

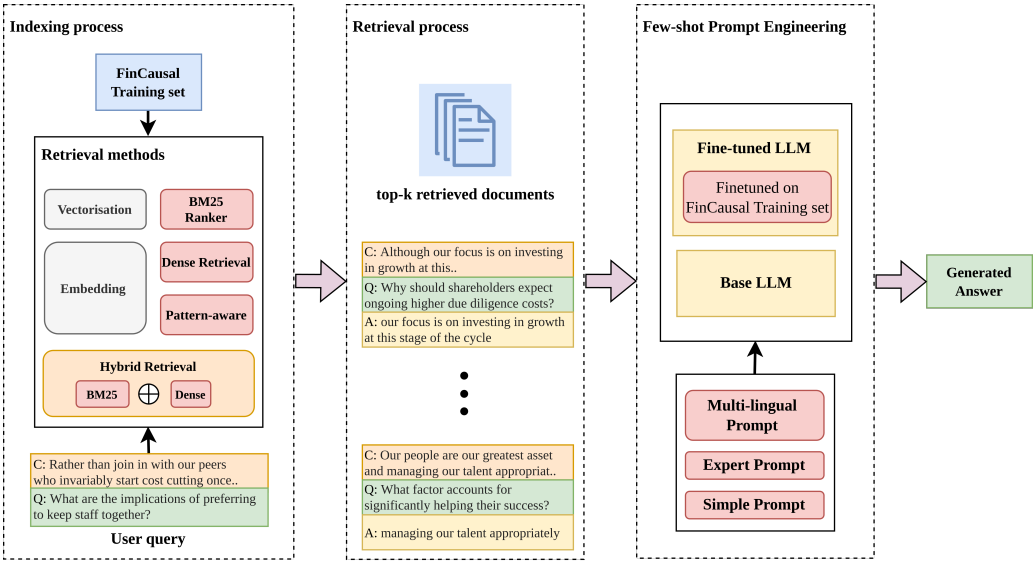


Figure 1: Overview of system architecture. The pipeline consists of three stages: (1) **Indexing**, where the training set is indexed using multiple retrieval methods (Random, RAG-BM25, RAG-Dense, RAG-Pattern, and RAG-Hybrid); (2) **Retrieval**, where the top-k most relevant training examples are retrieved for each test query; and (3) **Few-shot Prompt Engineering**, where the retrieved examples are combined with the test query and passed to either the base or fine-tuned LLM with different prompt strategies to generate the answer.

4.2. Prompt Engineering

Prompt engineering was employed to guide the GPT model to answer causal questions from the provided passages. The model was configured with a temperature of 0.1 and top-p of 1 to ensure deterministic outputs across multiple runs, and a maximum token limit of 512. We began with zero-shot prompting and progressively refined our approach using few-shot examples from the training set. To meet the task's extractive requirements, we implemented strict instructional constraints: the model was explicitly directed to answer the given question by extracting the answer verbatim from the context C and it was strictly prohibited from paraphrasing.

Three prompting strategies were explored, as illustrated in Figure 1: (1) a *simple prompt*, which provides minimal instructions to extract the answer from the context; (2) an *expert prompt*, which assigns the model a domain-expert role (e.g., "You are a financial causal analysis expert") and includes detailed extraction rules; and (3) a *multilingual prompt*, which extends the expert prompt with language-aware instructions to handle both English and Spanish inputs (e.g., "The passage and question may be in Spanish.").[1]

Two strategies were explored for selecting few-shot examples:

1) Random sampling. We sampled k examples randomly from the training set. The model was provided with k question-context-answer triplets as examples before being asked to answer the target question. We experimented with $k \in \{5, 10\}$.

2) RAG-based approaches. The most relevant training examples for each test instance were retrieved using the RAG approach (Lewis et al., 2020). Multiple retrieval strategies were explored: (1) BM25 (Robertson and Zaragoza, 2009), a sparse lexical retrieval method widely used for relevance ranking in information retrieval (RAG-BM25); (2) a dense semantic retrieval approach that encodes texts into dense vector representations and retrieves the most semantically similar training examples based on vector similarity (RAG-Dense) (Karpukhin et al., 2020); (3) a pattern-aware variant (RAG-Pattern), in which each question is first classified into a causal template (CAUSE, EFFECT, or OTHER), and retrieval is restricted to examples of the same type; and (4) a hybrid approach that combines RAG-BM25 and RAG-Dense using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to merge the ranked lists into a single unified ranking (RAG-Hybrid).

For RAG-BM25, every question–context pair in the training dataset was indexed using a BM25 retriever with PyStemmer-based language-specific stemming for both English and Spanish.[2] For RAG-Dense, embeddings were generated for the same training examples using OpenAI's `text-`

[1] Prompt templates: https://github.com/aaalqarni/fincausal-2026/blob/main/prompts/README.md

[2] https://pypi.org/project/PyStemmer/

`embedding-3-large`[3] model and indexed using LlamaIndex's in-memory vector store (Liu, 2022). For RAG-Pattern, each question was first classified into the causal class using regular expression patterns, and retrieval was restricted to training examples of the same template class. At inference time, each test question–context pair was embedded using the same encoder, and the top-k most similar training examples were retrieved and provided as few-shot examples to the base and the fine-tuned GPT-4.1-mini models.

4.3. Fine-tuning

Following the approach of the top-performing teams in FinCausal 2025 (Moreno Sandoval et al., 2025), GPT-4.1-mini was fine-tuned on the provided training data using OpenAI's fine-tuning API. The training set was formatted as question-context-answer in JSONL format, where each entry contained the question, context, and expected extractive answer. The training process ran for 3 epochs with a learning rate multiplier of 2 and a batch size of 3. This configuration was specifically chosen to accelerate convergence and to avoid overfitting, given the limited size of the dataset.[4]

5. Experimental Setup

Experimental Stages. Experiments were conducted in three different stages. First, various prompt templates and few-shot configurations ($k \in \{0, 5, 10\}$) were evaluated on the development set using the base GPT-4.1-mini model. Second, GPT-4.1-mini was fine-tuned on the 3,200 training instances and its performance was benchmarked against the base version across all retrieval strategies. Third, ablation studies were conducted on the five retrieval methods described in Section 4, varying the number of few-shot examples. In the hybrid configuration, RRF was used with equal weighting between the lexical (RAG-BM25) and semantic (RAG-Dense) retrieval examples.

Final Submission. For the final submission, the retrieval indexes were rebuilt using all 2,000 training instances per language, and predictions were generated on the blind test sets using the best-performing configuration from the development experiments.

Evaluation Metrics. EM and SAS (Risch et al., 2021) are reported for the development set. EM measures whether the predicted answer exactly matches the gold answer string. SAS measures the semantic similarity between the predicted and gold answers using a cross-encoder model. For SAS, `all-MiniLM-L6-v2` is used for English and `paraphrase-multilingual-MiniLM-L12-v2` for Spanish. The official competition ranking uses LLM-as-a-judge scoring on the blind test set, which rates system outputs on a 1–5 adequacy scale.

[3] https://developers.openai.com/api/docs/models/text-embedding-3-large

[4] https://developers.openai.com/api/docs/guides/model-optimization

6. Results and Discussion

For English, the simple prompt performed best, while the multilingual prompt produced the best results for Spanish. Table 2 presents our results across all configurations for both English and Spanish sub-tasks. Results showed a large performance gap between the base and fine-tuned models. Fine-tuning GPT-4.1-mini on bilingual (EN+ES) data increased English EM from .3533 to .8875 and Spanish EM from .0250 to .8625. This improvement was consistent across all retrieval configurations, suggesting that fine-tuning helps the model follow the strict verbatim extraction requirement rather than relying only on in-context examples.

For the fine-tuned model, all retrieval strategies produced comparable results. English EM ranged from .8650 to .8875, and Spanish EM from .8425 to .8625. The difference between the best configuration RAG-Dense (k=5) and the worst was below 2 percentage points in both languages. In contrast, the base model benefited more from retrieval. For example, RAG-Hybrid (k=10) achieved .5950 EM in English compared to .3533 in the zero-shot setting. This suggests that retrieval approaches play a larger role when the model is not fine-tuned, whereas fine-tuning reduces the additional gains from complex retrieval strategies. RAG-Pattern showed higher EM for the base model in English (RAG-Pattern k=10: .7550 vs RAG-Dense k=10: .5875), but underperformed on SAS and showed no consistent gains for the fine-tuned model, suggesting that fine-tuning implicitly captures causal directionality.

RAG benefits (for Spanish). The base model showed lower zero-shot performance on Spanish (EM = .0250) compared to English (EM = .3533), likely due to a higher prevalence of English financial corpora in the pre-training data. However, RAG strategies narrowed this gap. Specifically, RAG-Dense (k=10) increased the Spanish EM score to .5075, a substantial relative improvement over the zero-shot baseline. This suggests that RAG few-shot examples are particularly beneficial in lower-resource scenarios.

Official Blind Test Results. On the official blind test set, evaluated using LLM-as-a-judge, the fine-tuned model with RAG-Dense (k=5) achieved the highest English score (4.8140), while RAG-Dense (k=10) obtained the best Spanish score (4.8131).

			English			Spanish		
			Validation		Blind	Validation		Blind
Model	**Mode**	k	**EM**	**SAS**	**LLM Score**	**EM**	**SAS**	**LLM Score**
GPT-4.1-mini	Zero-shot	0	.3533	.8605	4.4600	.0250	.8876	4.3002
	Random	5	.5450	.9083	4.4620	.3450	.9282	4.3857
		10	.5850	**.9354**	4.4760	.3925	.9383	4.4076
	RAG-BM25	5	.5675	.9113	4.4920	.4400	.9393	4.4513
		10	.5800	.9135	4.5280	.4850	.9395	4.5189
	RAG-Dense	5	.5550	.9136	4.5780	.4525	.9347	4.4712
		10	.5875	.9151	4.5840	**.5075**	.9408	**4.5030**
	RAG-Pattern	5	.7300	.8593	4.6220	.4075	.9406	4.4135
		10	**.7550**	.8622	**4.6500**	.4175	**.9432**	4.4334
	RAG-Hybrid	10	.5950	.9152	4.5940	.4575	.9372	4.4712
Finetuned GPT-4.1-mini (EN+ES)	Zero-shot	0	.8800	.9755	4.7440	.8550	.9734	4.7853
	Random	5	.8800	.9763	4.7940	.8600	.9735	4.7952
		10	.8675	.9776	4.7860	.8575	.9732	4.8012
	RAG-BM25	5	.8775	.9741	4.7980	.8550	.9740	4.7932
		10	.8825	.9770	4.7920	.8475	.9734	4.8012
	RAG-Dense	5	**.8875**	**.9790**	**4.8140**	**.8625**	.9720	4.8032
		10	.8825	.9759	4.7940	.8425	.9698	**4.8131**
	RAG-Pattern	5	.8650	.8796	4.7380	.8500	**.9780**	4.7714
		10	.8700	.8801	4.7080	.8550	.9780	4.7773
	RAG-Hybrid	10	.8850	.9773	4.7980	.8600	.9724	4.7932

Table 2: Results across English and Spanish sub-tasks. EM (Exact Match) and SAS (Semantic Answer Similarity) are computed on the 400-instance validation set per language. LLM denotes the official blind test score evaluated by an LLM-as-judge on a 1–5 scale. Bold values indicate the best result per metric within each model group.

Model	**Mode**	**k**	**English**	**Spanish**
Finetuned GPT-4.1-mini	RAG-Dense	(5, 10)	**4.8140**	**4.8131**
Our Ranking			*1st (tie)*	*1st*

Table 3: Official rankings based on LLM scores (out of 5) on the blind test set for the FinCausal 2026 shared task. The system used 5-shot prompting for English and 10-shot prompting for Spanish.

These configurations were ranked first among all participating systems for both sub-tasks, as illustrated in Table 3. A small variance was observed across fine-tuned settings (4.7440–4.8140 for English; 4.7853–4.8131 for Spanish), indicating stable performance across different RAG approaches. These results demonstrate that fine-tuned GPT-4.1-mini combined with RAG-Dense is the most effective approach for causal Q&A in financial texts.

Error Analysis. Error analysis was conducted on the development set, as gold answers for the official test set were not released to participants. We analysed the mismatched predictions from the best-performing fine-tuned configuration RAG-Dense (k=5) and identified two main types of errors: (1) *span boundary errors*, where the model extracted a slightly longer or shorter span than the gold answer, typically by including or omitting a leading clause; and (2) *causal direction confusion*, where the model confused the direction of causality when multiple cause–effect relations were present in the context. For instance:

Q: "Why did Legendary increase its stake in Virtual Stock from 6.8% to 7.2%?"

C: "Subsequent to this investment and also in July 2017, Legendary increased its stake in Virtual Stock from 6.8% (subsequent to **the dilution due to Notion's investment**) to 7.2% increasing the carrying value of its investment in Virtual Stock to £4.3m." (A indicated by **bold font**)

A (predicted): *increasing the carrying value of its investment in Virtual Stock to £4.3m*.

These errors highlight the difficulty of distinguishing between causal explanations and subsequent financial outcomes in complex financial narratives.

7. Conclusion and Future Work

This paper presented a systematic comparison of RAG approaches combined with fine-tuning for causal Q&A in financial texts in English and Spanish. Fine-tuning GPT-4.1-mini on bilingual data improved performance across both languages. RAG-Dense further improved performance, achieving the best results in both sub-tasks. RAG-Pattern improved base model performance but showed no consistent gains after fine-tuning, suggesting that fine-tuning reduced reliance on causal pattern matching. Our system ranked first in both sub-tasks. Future work will explore more advanced RAG approaches and extend the approach to other domains and languages.

Code Availability

Code to reproduce experimental results reported in this paper is publicly available[5].

Bibliographical References

Bushra Aljohani and Tawfeeq Alsanoosy. 2026. Enhancing medical question answering with llms via a hybrid retrieval-augmented generation framework. *Information*, 17(2):133.

Lang Cao, Shihua Zhang, and Juxing Chen. 2022. Cbcp: A method of causality extraction from unstructured financial text. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '21, page 135–140, New York, NY, USA. Association for Computing Machinery.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ying Li, Xiaosha Xue, Zhipeng Liu, Peibo Duan, and Bin Zhang. 2024. Implicit-causality-exploration-enabled graph neural network for stock prediction. *Information*, 15:743.

Jerry Liu. 2022. LlamaIndex.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stéphane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020).

Antonio Moreno Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The financial document causality detection shared task (fincausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The

[5] https://github.com/aaalqarni/fincausal-2026

financial document causality detection shared task (fincausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination in causal Q&A: The efficacy of fine-tuning over prompting in LLMs. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 253–258, Abu Dhabi, UAE. Association for Computational Linguistics.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Neelesh K Shukla, Raghu Katikeri, Msp Raja, Gowtham Sivam, Shlok Yadav, Amit Vaid, and Shreenivas Prabhakararao. 2023. Investigating large language models for financial causality detection in multilingual setup. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2866–2871. IEEE.

Tianyi Yang, Nashrah Haque, Vaishnave Jonnalagadda, Yuya Jeremy Ong, Zhehui Chen, Yanzhao Wu, Lei Yu, Divyesh Jadav, and Wenqi Wei. 2026. Augmenting question answering with a hybrid rag approach.

Language Resource References

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Moreno-Sandoval, Antonio and Torterolo Orta, Yanco Amor and Stanescu, Maria Alexia and Chatzi, Melina. 2026. *The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset*. e-cienciaDatos.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

Causal Connections: Leveraging Multilingual Fine-Tuning for Financial QA@FinCausal 2026

Akash Kumar Gautam, Serhii Hamotskyi, Christian Hänig
Anhalt University of Applied Sciences
{akash-kumar.gautam, serhii.hamotskyi, christian.haenig}@hs-anhalt.de

Abstract

This paper describes team HSA_CORAL's submission to the FinCausal 2026 shared task on extracting cause–effect relations from financial narratives via extractive question answering in English and Spanish. We compare three modeling families: (i) encoder-only token tagging with multilingual BERT, (ii) encoder–decoder generation with multilingual BART, and (iii) decoder-only LLMs (Llama 3.1 and GPT variants) using prompt refinement, few-shot demonstrations, and supervised fine-tuning. Across settings, prompting and few-shot examples yield competitive performance, while supervised fine-tuning provides the largest gains. Our best system, GPT-4.1 Mini fine-tuned on combined English and Spanish training data, achieves a tied highest score on the English subtask (score 4.8140) and ranks third on Spanish (score 4.7753) under the shared task's LLM-as-a-judge metric. Overall, the results highlight the value of task-specific adaptation and multilingual fine-tuning for cross-lingual transfer in financial causality QA.

Keywords: Large Language Models, Financial NLP, Financial Question Answering.

1. Introduction

Understanding cause–effect relationships in financial texts is essential for informed decision-making. They reveal drivers of stock prices, economic changes, market behavior, and regulatory decisions across different countries. For analysts and investors, identifying causal connections provides valuable insights into financial risks, potential investments, and strategic planning (Gopalakrishnan et al., 2023).

The FinCausal shared task has steadily advanced how we detect causality in finance. Earlier editions focused on identifying causal phrases directly in text. Later versions introduced more complex challenges, such as recognizing implicit causality and multi-step reasoning (Mariko et al., 2022, 2020; Zavitsanos et al., 2023). The most recent task required generative models to answer open-ended questions about causes and effects, using Exact Match (EM) and Semantic Answer Similarity (SAS) (Risch et al., 2021) as evaluation metrics (Moreno-Sandoval et al., 2025).

The 2026 edition introduces several new elements (Moreno-Sandoval et al., 2026). The dataset includes fragments with new complex causal relationships, rephrased questions that demand more sophisticated reasoning, and randomly divided train-test splits. A novel challenge is the use of an LLM-as-a-judge for evaluation, which rates responses on a 1–5 adequacy scale. The task pushes models beyond simple text matching toward understanding both explicit and implicit causal relationships in detailed financial narratives.

We explored three approaches to extractive question-answering: (i) token classification with encoder-only models like BERT, (ii) sequence-to-sequence generation with encoder-decoder models like BART, and (iii) few-shot prompting with decoder-only large language models.

Across our experiments, we find that fine-tuned generative models consistently outperform encoder-only and encoder-decoder architectures on the extractive QA task. Notably, GPT-4.1 Mini fine-tuned on a multilingual corpus achieves the best overall performance, securing the top position on the English subtask and third place on Spanish. Adding 20 few-shot examples proves critical, enabling even a smaller model to surpass its larger counterparts in zero-shot settings. These results demonstrate that multilingual fine-tuning and few-shot learning together provide a reliable approach for causal relation extraction in financial documents.

2. Related Work

Causal Information Extraction Early approaches to causal relationship detection in financial texts relied heavily on rule-based systems and traditional machine learning methods such as Support Vector Machines (SVMs) and decision trees (Ghosh and Naskar, 2022; Baranes et al., 2019). While these models demonstrated some success in identifying patterns in financial reports and news articles, they required extensive feature engineering and often struggled to capture the complexity and temporal dynamics of causal relations in domain-specific documents.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 132–138

The introduction of BERT (Devlin et al., 2019) and its multilingual variants marked a significant shift in the field (Yang et al., 2019; Wan and Li, 2022). These pre-trained language models enabled more nuanced understanding of contextual information with minimal feature extraction, substantially improving performance on tasks involving document-level comprehension (Zhang and Jankowski, 2022).

Subsequent work has focused on fine-tuning these models on domain-specific financial datasets, achieving state-of-the-art results in tasks such as sentiment analysis and event extraction (Mariko et al., 2020). Fine-tuned pre-trained language models have consistently outperformed traditional machine learning approaches, particularly when working with large-scale financial reports (Jin et al., 2023; Huang et al., 2023; Sarmah et al., 2023). Liu et al. (2023) propose an implicit cause–effect interaction framework to improve event causality extraction.

Pretrained Language Models More recently, proprietary large language models such as GPT-4 have been explored for question-answering tasks in the financial domain (Zhang et al., 2023; Kalpakchi and Boye, 2023). These models have demonstrated strong few-shot learning capabilities, enabling them to generalize from limited examples without task-specific fine-tuning (Xiao et al., 2022; Guo et al., 2023). This line of work underscores the growing potential of generative models for specialized NLP tasks, including the extraction of causal relationships in finance (Nayak et al., 2022; Kim et al., 2023).

3. Methodology

We compare three approaches to extractive QA for FinCausal 2026: (1) token classification using BERT-based models, (2) extractive QA using encoder-decoder models, and (3) generative models. These approaches are compared across a variety of pretrained large language models in few-shot settings using a multilingual dataset.

3.1. Encoder-Based Extractive QA

This approach utilizes text embedding models such as BERT for token classification, following a similar methodology to Yoon et al. (2022). The process begins by tokenizing both the context and the question, which are then concatenated with a special `[SEP]` token. During training, each sample is annotated using an IO tagging scheme, where the answer span is mapped to its first occurrence within the passage.

We compute the cross-entropy loss between the predicted token classes and the ground truth labels derived from the training data. To refine the loss calculation, a loss mask is applied to restrict the computation exclusively to tokens originating from the passage. This masking strategy excludes tokens from the question and special symbols (e.g., `[SEP]`, padding). This focuses learning on the passage tokens.

3.2. Encoder-Decoder Extractive QA

Our second approach treats extractive QA as a sequence-to-sequence generation task using BART (Lewis et al., 2020). The input consists of the question and context concatenated with a delimiter, and the model is trained to generate the answer token-by-token.

We fine-tune BART by minimizing the negative log-likelihood of the target answer sequence. During inference, beam search is employed to decode the most likely answer. This formulation is effective for extractive tasks as it leverages BART's pretrained language understanding while flexibly handling answer spans. We evaluate multilingual BART variants on both the English and Spanish subtasks.

3.3. Decoder-Based Extractive QA

The flexibility of large language models makes them well-suited for extractive QA tasks. However, it is important to ensure that models follow instructions and avoid hallucinations beyond the provided context (Xu et al., 2024). To address this, we implement a multi-step strategy combining prompt optimization, few-shot selection, and fine-tuning.

First, we perform prompt optimization through several iterative refinements on a small subset of the training dataset. The final version of the prompt used is described in Appendix A. This process helps identify instruction formats and phrasings that consistently yield extractive, context-grounded answers. We incorporate few-shot examples within the prompt, selecting relevant QA demonstrations using cosine similarity. Specifically, given a test context and question, we retrieve the most similar QA pairs from the training set based on multilingual embedding similarity[1], and include them as examples in the prompt to guide the model's output.

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Model	Fine-tuned	Corpus	English	Spanish
BERT Base Multilingual	yes	Multilingual (en+es)	3.9800	3.9810
BART Facebook Base	yes	Multilingual (en+es)	4.1200	4.0300
Llama-3.1 8B	yes	Multilingual (en+es)	4.0200	3.9100
GPT-3.5 turbo	no	-	4.7040	4.7060
GPT-4.1 mini	yes	Monolingual (en)	4.7560	4.7141
GPT-4.1 mini	yes	Monolingual (es)	4.7210	4.7674
GPT-4.1 mini	yes	Multilingual (en+es)	**4.8140**	**4.7753**
GPT-5.2	no	-	4.7600	4.7350

Table 1: Evaluation results on blinded English and Spanish test sets, with scores provided by an external LLM judge. The *Corpus* column indicates whether models were fine-tuned on monolingual or multilingual data. Bold entries denote the highest score in each language column. Decoder-only models were tested with varying numbers of few-shot examples; results, obtained with 20 examples, are reported here. Best results using our approach are presented in **bold**.

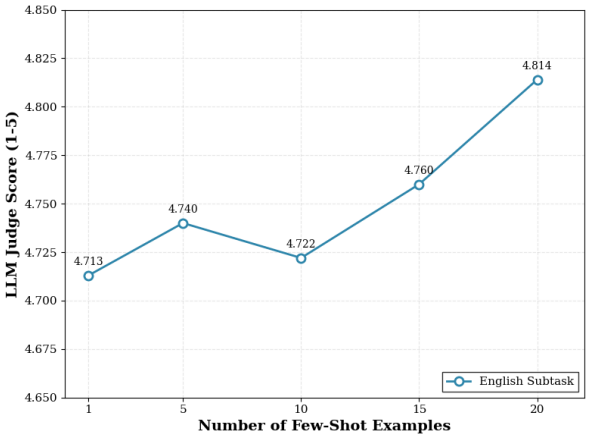


Figure 1: External LLM-judge score (English subtask) as a function of the number of few-shot demonstrations in the prompt for the best-performing decoder-only configuration.

As a further enhancement, we fine-tune the model on up to 2,000 samples, experimenting with three configurations: training on English data only, Spanish data only, and a combined bilingual dataset. Fine-tuning reinforces task-specific behavior and significantly reduces hallucinations, leading to more reliable extractive answers.

This hybrid approach—combining prompt engineering, dynamic few-shot selection, and targeted fine-tuning—allows us to leverage the generative strength of LLMs while maintaining the extractive precision required for the task.

4. Experiments

4.1. Dataset

The dataset comprises financial narratives in English and Spanish designed for an extractive question answering task focused on causal relationships Moreno-Sandoval et al. (2026). Given a context and a question, the task requires identifying a text span that expresses a causal relation within the financial narrative (Moreno-Sandoval et al., 2025).

Questions are formulated abstractly, targeting either the cause or the effect described in the text, with causes defined as agents or facts that can be extracted verbatim from the provided context.

For complex causal structures—such as causal chains or non-linear relationships—up to two questions are included per context to ensure comprehensive coverage. The English dataset is sourced from financial annual reports from 2017, drawn from the UCREL corpus and the English portion of the 2018 FinT-esp corpus, while the Spanish dataset is extracted from a corpus of Spanish financial annual reports spanning 2014 to 2018. The training set for both subtasks consists of 2,000 samples, with test sets of 500 samples for English and 503 samples for Spanish, respectively.

4.2. Model Selection

BERT[2] was used as an encoder model, multilingual BART[3] for encoder-decoder model and we evaluated Llama-3.1 and several GPT[4] variants for

[2]https://huggingface.co/google-bert/bert-base-multilingual-cased

[3]https://huggingface.co/facebook/bart-base

[4]https://developers.openai.com/api/docs/models

decoder-only setups. For fine-tuning of Llama3.1 we used Low-Rank Adaptation (LoRA) to speed the process (Hu et al., 2022).

4.3. Results

Table 1 reports the best-performing configuration for each model family. Scores correspond to the shared task's external LLM-as-a-judge evaluation on the blind test set, rated on a 1–5 adequacy scale for both English and Spanish.

Encoder and Encoder-Decoder Setups. Extractive models such as BERT perform reasonably well at identifying exact spans corresponding to causal relationships, particularly when the answer is explicitly stated in the context. However, BART consistently outperforms BERT across both languages, benefiting from its sequence-to-sequence formulation which allows for more flexible and accurate span generation. Interestingly, the multilingual variant of BERT fine-tuned on combined English and Spanish data yields better performance on the blinded test set than its monolingual counterparts trained on individual languages. For brevity, we report only the best-performing configurations.

Few-Shot Learning with Decoder-Only Models In our decoder-only experiments, we evaluated Llama 3.1 and several variants of GPT. For fine-tuning Llama 3.1, we employed Low-Rank Adaptation (LoRA) to reduce computational overhead while maintaining task performance. Our findings indicate that while well-structured prompts with clear instructions are effective at reducing hallucinations, the inclusion of few-shot examples substantially improves the quality and precision of the generated answer spans. Notably, adding examples from the training set resulted in substantial improvements for GPT-4.1 Mini compared to zero-shot settings—even enabling it to outperform a much larger model (GPT-5.2[5]) on both Spanish and English subtasks.

Figure 1 plots the LLM-as-a-judge scores against the increasing number of few-shot examples included in the prompt. The trend shows a clear improvement in output quality as more examples are added, with scores increasing correspondingly on the leaderboard. However, increasing the number of examples beyond a certain point did not yield further gains; in some cases, it even led to the generation of hallucinated content.

Given the context window constraints of each model, we systematically varied the number of few-shot examples. The best results across all decoder-only models were achieved with 20 few-shot examples, which we adopt as our final configuration.

Fine-Tuning Our experiments reveal a clear advantage of fine-tuned generative models over both encoder-only and encoder-decoder architectures. Multilingual fine-tuning consistently outperforms monolingual fine-tuning on both English and Spanish subtasks, demonstrating strong cross-lingual transfer for causal relation extraction in financial narratives. Among all configurations, GPT-4.1 Mini with multilingual fine-tuning achieved the best results across both languages.

Notably, this fine-tuned model outperforms a much larger model, GPT-5.2, in zero-shot settings. This suggests that the test dataset's contexts and questions contain lexical characteristics and causal relationships (for both Spanish and English) that can be reliably identified only when models have access to task-specific annotation guidelines through fine-tuning. The observation points to an interesting conclusion: fine-tuned pre-trained language models of moderate size may offer better performance for specialized tasks than much larger models used in inference-only mode, highlighting the importance of fine-tuning over parameter scaling for domain-specific applications.

Error Analysis and Limitations While our models achieved strong overall results, error analysis reveals persistent errors in handling nested causal structures and contexts containing multiple potential causes. In these challenging cases, models often selected the wrong causal pair, with errors more frequent in the Spanish subtask—suggesting that cross-lingual generalization remains challenging for complex causal structures.

The lack of detailed annotation guidelines for identifying relevant spans further compounds this issue, as prompt optimization alone cannot fully resolve such ambiguities. Looking ahead, alignment techniques such as Direct Preference Optimization (Rafailov et al., 2023) and Reinforcement Learning from Human Feedback (Bai et al., 2022) offer promising ways for learning the implicit patterns underlying human annotations for structurally complex examples.

5. Conclusion

This work demonstrates that generative models outperform encoder and encoder-decoder architectures on causal question-answering tasks for

[5]OpenAI currently supports supervised fine-tuning for GPT models up to GPT-4.1-2025-04-14.

financial narrative documents, provided that hallucinations are mitigated through careful prompt engineering and few-shot examples. Multilingual fine-tuning with a large number of training samples from both Spanish and English, significantly boosts performance, highlighting the effectiveness of cross-lingual transfer in this domain. Future work may explore LLM-based evaluation as a means to further enhance output quality, as well as more sophisticated strategies for handling nested and ambiguous causal structures.

Acknowledgments

This work has been completed as part of project CORAL (Constrained Retrieval-Augmented Language Models), funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under grant number 16IS24077C.

6. References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Amos Baranes, Rimona Palas, et al. 2019. Earning movement prediction using machine learning-support vector machines (svm). *Journal of Management Information and Decision Sciences*, 22(2):36–53.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Sohom Ghosh and Sudip Kumar Naskar. 2022. Lipi at fincausal 2022: Mining causes and effects from financial texts. In *Proceedings of the 4th financial narrative processing workshop@ LREC2022*, pages 121–123.

Seethalakshmi Gopalakrishnan, Victor Zitian Chen, Wenwen Dou, Gus Hahn-Powell, Sreekar Nedunuri, and Wlodek Zadrozny. 2023. Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs. *Information*, 14(7):367.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Yiqiao Jin, Xiting Wang, Yaru Hao, Yizhou Sun, and Xing Xie. 2023. Prototypical fine-tuning: Towards robust performance under varying data sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12968–12976.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in swedish using gpt-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491.

Yuheun Kim, Lu Guo, Bei Yu, and Yingya Li. 2023. Can chatgpt understand causal language in science claims? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 379–389.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin, and Xiaoyu Li. 2023. Event causality extraction via implicit cause-effect interactions. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6792–6804.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality

extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 105–107.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (fincausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA.

Tapas Nayak, Soumya Sharma, Yash Butala, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2022. A generative approach for financial causality extraction. In *Companion Proceedings of the Web Conference 2022*, pages 576–578.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157.

Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, and Tianjie Zhu. 2023. Towards reducing hallucination in extracting information from financial reports using large language models. In *Proceedings of the Third International Conference on AI-ML Systems*, pages 1–5.

Chang-Xuan Wan and Bo Li. 2022. Financial causal sentence recognition based on bert-cnn text classification. *The Journal of Supercomputing*, 78(5):6503–6527.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11204–11213.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 72–77.

Wonjin Yoon, Richard Jackson, Aron Lagerberg, and Jaewoo Kang. 2022. Sequence tagging for biomedical extractive question answering. *Bioinformatics*, 38(15):3794–3801.

Elias Zavitsanos, Aris Kosmopoulos, George Giannakopoulos, Marina Litvak, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, and Mo El-Haj. 2023. The financial narrative summarisation shared task (fns 2023). In *2023 IEEE International Conference on Big Data, BigData 2023*, pages 2890–2896. Institute of Electrical and Electronics Engineers.

Le Zhang, Yihong Wu, Fengran Mo, Jian-Yun Nie, and Aishwarya Agrawal. 2023. Moqagpt: Zero-shot multi-modal open-domain question answering with large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1195–1210.

Ning Zhang and Maciej Jankowski. 2022. Hierarchical bert for medical document understanding. *arXiv preprint arXiv:2204.09600*.

7. Language Resource References

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

A. Prompt

Figure 2 describes the prompt used by decoder only language models for generating the final response as part of the extractive question-asnwering task.

Task Description
Given a financial context and a question, extract an exact answer from the context about cause or effect that addresses the question. The answer will be either the cause or the effect to a specific event mentioned in the context.

Instructions

1. **Read the context carefully:**
    - Understand the events and relationship described in the context.
2. **Understand the question:**
    - Determine if the question is asking for cause or effect.
    - Identify the specific events or statement the question refers to.
3. **Extract the answer verbatim:**
    - Locate the exact sentence or phrase in the context that answers the question.
    - **Do not paraphrase, summarise, or add any external information. The answer must be copied word-for-word from the context.**
4. **Provide only the answer:**
    - **Do not include any instructions, explanations or formatting.**
    - **Output only the extracted answer and nothing else.**
5. **Answer should be in the language in which question is asked and the context is mentioned:**
    - **Understand the context very well.**

Examples
`{ { formatted_examples } }`

Your Task
Context:
`{ { context } }`
Question:
`{ { question } }`
Answer:
[Provide only the exact answer from the context]

Remember

- **Output only the answer. Do not include any additional text. Do not include "Answer" in your output.**
- **The answer must exactly match a portion of the context.**
- **Do not add instructions, explanations, or any extra information.**

Figure 2: Prompt used for extractive QA by decoder models used in the FinCausal 2026 shared task.

VERSA: Verbatim Extraction via Rephrasing and Self-Aggregation for Financial Causality

Aldan Jay, Rafael Berlanga, Yoelvis Moreno, Vicent Santamarta
Escola de Doctorat, Universitat Jaume I, Castellón de la Plana, Spain
Universitat Jaume I, Castellón de la Plana, Spain
{jay, berlanga, alcayde, santamav}@uji.es

Abstract

Financial causality detection, the task of identifying and extracting verbatim causal spans from financial narratives, remains a challenging problem in Natural Language Processing (NLP). Large Language Models (LLMs), while powerful reasoners, frequently paraphrase source text or produce imprecise span boundaries when used in zero-shot extraction settings, leading to poor Exact Match scores. In this paper, we present VERSA, our system for the FinCausal 2026 Shared Task, a multi-agent pipeline that integrates two complementary inference strategies: Rephrase-and-Respond (RaR) and Recursive Self-Aggregation (RSA). The pipeline decomposes the extraction task into five sequential stages, each handled by a specialised agent: (1) causal structure analysis, (2) question reformulation via RaR, (3) diverse candidate population generation, (4) iterative refinement through RSA, and (5) verbatim validation with word-boundary alignment. We evaluate our approach on both the English and Spanish subsets of the FinCausal 2026 dataset. An ablation study demonstrates the individual and combined contributions of RaR and RSA, showing that the full pipeline substantially outperforms a zero-shot baseline in Exact Match and token-level F1.

Keywords: financial causality, extractive question answering, multi-agent systems, large language models, recursive self-aggregation

1. Introduction

The automatic extraction of causal relationships from financial documents is a long-standing challenge in Financial NLP. Causal reasoning is central to financial analysis: understanding why revenue declined, what drove a loss, or which factors contributed to market movements requires precise identification of cause–effect spans within narrative text. The FinCausal shared task series (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023; Moreno Sandoval et al., 2025) has established a rigorous evaluation framework for this problem, requiring systems to return *verbatim* text spans that correctly identify the cause or effect queried in a given question.

Modern Large Language Models (LLMs) have demonstrated strong performance in a wide range of NLP tasks (Brown et al., 2020), including question answering and information extraction. However, when applied to extractive tasks requiring exact span matching, LLMs exhibit systematic weaknesses. In zero-shot settings, they tend to *paraphrase* rather than extract, they *hallucinate* causal connectives (e.g., prepending "due to" to the extracted span), and they produce *inconsistent span boundaries*—for instance, omitting an initial determiner or including trailing punctuation. These behaviours, while often semantically harmless, are severely penalised by Exact Match (EM) metrics.

The 2026 edition of FinCausal (Moreno-Sandoval et al., 2026) introduces additional complexity: over 500 new fragments containing multi-element causal chains, abstractive question rephrasing in approximately 10% of the dataset, and a new LLM-as-a-judge evaluation metric scored on a 1–5 adequacy scale (Zheng et al., 2024). These changes demand systems capable of deep reasoning beyond surface-level lexical matching.

To address these challenges, we propose **VERSA** (***V**erbatim **E**xtraction via **R**ephrasing and **S**elf-**A**ggregation*), a multi-agent pipeline that decomposes the extraction task into five specialised stages. Our approach integrates two key techniques from recent research:

1. **Rephrase-and-Respond (RaR)** (Deng et al., 2023): a prompting strategy in which the model reformulates the input question to align it with its own internal reasoning frame, thereby reducing ambiguity prior to extraction.

2. **Recursive Self-Aggregation (RSA)** (Venkatraman et al., 2025): an iterative refinement mechanism that maintains a population of candidate answers and recursively aggregates subsets to converge towards a robust consensus, rather than relying on a single inference pass or simple majority voting.

VERSA operates on both the English and Spanish portions of the FinCausal 2026 dataset. The remainder of this paper is organised as follows: Section 2 reviews related work; Section 3 describes our methodology in detail; Section 4 presents the experimental setup; Section 5 reports results and

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 139–145

an ablation study; and Section 6 offers concluding remarks.

2. Related Work

2.1. Financial Causality Detection

The FinCausal shared task series, initiated at COLING 2020 (Mariko et al., 2020), formulated financial causality detection as an extractive question answering (QA) problem. Subsequent editions (Mariko et al., 2022; Moreno-Sandoval et al., 2023; Moreno Sandoval et al., 2025) refined the annotation scheme and expanded coverage to Spanish and multi-element causal chains, culminating in the 2026 edition (Moreno-Sandoval et al., 2026) evaluated in this work. Given a financial text passage and a causal question, systems must return the exact text span containing the queried cause or effect. Previous editions have seen strong participation from systems based on fine-tuned Transformer encoders such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), often coupled with Conditional Random Fields (CRFs) for token-level sequence labelling (Lafferty et al., 2001). While effective, these approaches require task-specific fine-tuning and struggle with complex, multi-hop causal chains where the answer spans multiple clauses.

2.2. Prompting Strategies for Extraction

The advent of instruction-tuned LLMs has enabled zero-shot and few-shot extractive QA without task-specific training (Brown et al., 2020). However, LLMs frequently reformulate rather than extract, a fundamental tension between their generative nature and the extractive requirement of tasks like FinCausal. Chain-of-Thought prompting (Wei et al., 2022) and Self-Consistency (Wang et al., 2023) have improved reasoning quality, but neither directly addresses the verbatim constraint. The Rephrase-and-Respond (RaR) framework (Deng et al., 2023) proposes that LLMs reformulate ambiguous questions in their own terms before answering, effectively aligning the query with the model's internal processing frame. We adapt this insight specifically for extractive QA, using the rephrased question to generate explicit extraction hints.

2.3. Aggregation-Based Inference

Recursive Self-Aggregation (RSA) (Venkatraman et al., 2025) extends the self-consistency paradigm by maintaining a population of N candidate solutions and iteratively recombining randomly sampled subsets of size K over T iterations. Unlike majority voting, which selects the most frequent answer, RSA synthesises new candidates from subsets, enabling the correction of partial errors and convergence towards more complete answers. This approach has demonstrated improvements in mathematical reasoning and code generation, but has not, to our knowledge, been applied to extractive information extraction tasks.

3. Methodology

We propose a sequential multi-agent pipeline comprising five specialised agents, each responsible for a distinct stage of the extraction process. Figure 1 illustrates the overall workflow.

3.1. Stage 1: Causal Structure Analysis

The first agent analyses the input pair (c, q)—where c denotes the financial context and q the causal question—to produce a structured analysis $\mathcal{A}$ that guides subsequent stages. This analysis comprises three components:

Trigger detection. The agent identifies explicit causal markers in the context using pattern matching over language-specific lexicons. For English, these include expressions such as "due to," "as a result of," "driven by," and "consequently." For Spanish: "debido a," "como consecuencia de," "gracias a," and "por lo que." Each detected trigger is classified as *causal* (indicating a cause), *resultative* (indicating an effect), or *temporal-causal* (e.g., "following," "tras").

Directionality inference. The agent determines whether the question seeks the *cause* or the *effect* of the described relationship, a distinction critical for selecting the correct span when both are present in the context.

Complexity assessment. The number and arrangement of causal triggers are used to classify the relationship as *simple* (single cause–effect pair), *chain* (sequential cascade), or *multiple* (several concurrent causes or effects).

3.2. Stage 2: Question Reformulation (RaR)

Following Deng et al. (2023), who demonstrated that LLMs perform better when questions are restated in terms aligned with the model's internal reasoning, we apply a Rephrase-and-Respond step. Rather than forwarding the original question q directly to the extraction stage, the Rephraser agent generates a reformulated question q' and a set of extraction hints H.

The reformulation incorporates the causal analysis $\mathcal{A}$: it makes the target direction explicit (e.g., transforming "Why did X happen?" into "Identify the cause of X,") names specific entities from the context, and specifies the expected syntactic form

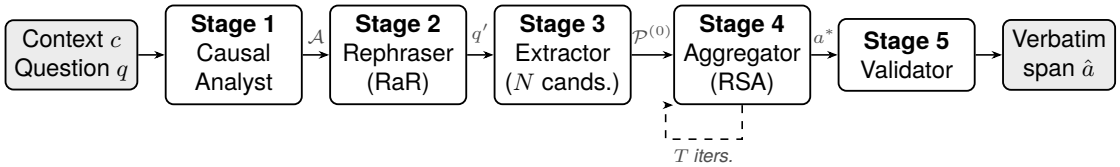


Figure 1: Overview of the proposed multi-agent pipeline. Stages 1–5 are executed sequentially. Stage 4 (Aggregator) performs T recursive iterations over the candidate population. Notation: $\mathcal{A}$ = causal analysis metadata; q' = rephrased question; $\mathcal{P}^{(0)}$ = initial candidate population; a^* = selected answer; $\hat{a}$ = final verbatim span.

of the answer (e.g., "a noun phrase following the marker 'due to."') This step reduces the ambiguity inherent in abstractive questions—particularly relevant for the 10% of FinCausal 2026 questions that have been deliberately rephrased away from the source text.

3.3. Stage 3: Candidate Population Generation

Rather than producing a single extraction, the Extractor agent generates a *population* $\mathcal{P}^{(0)} = \{a_1, a_2, \ldots, a_N\}$ of N candidate spans. This design choice is motivated by the observation that individual LLM inferences are stochastic: varying the decoding temperature or prompt formulation yields different, partially overlapping spans whose union often contains the correct answer.

The population is composed using two complementary methods:

- **LLM-based extraction**: The majority of candidates ($\lceil 3N/4 \rceil$) are obtained by querying the language model with the reformulated question q' and hints H at varying temperature values $\tau \in \{0.3, 0.5, 0.7, 1.0\}$. Each inference is independent, promoting diversity.
- **Machine Reading Comprehension (MRC)**: The remaining candidates ($\lfloor N/4 \rfloor$) are produced by a pre-trained multilingual extractive QA model (XLM-RoBERTa fine-tuned on SQuAD 2.0; Rajpurkar et al., 2018; Conneau et al., 2020), providing a non-generative anchor that is inherently verbatim.

Exact duplicate spans are removed to ensure diversity. In our experiments, we set $N = 8$.

3.4. Stage 4: Recursive Self-Aggregation (RSA)

The core refinement mechanism of our pipeline applies RSA (Venkatraman et al., 2025) to the candidate population. At each iteration $t \in \{1, \ldots, T\}$, we construct a new population $\mathcal{P}^{(t)}$ from $\mathcal{P}^{(t-1)}$ as follows:

1. For each position $i \in \{1, \ldots, N\}$, sample a subset $S_i \subset \mathcal{P}^{(t-1)}$ of size K uniformly at random without replacement.
2. Present the K candidates in S_i, together with the context c and the original question q, to the Aggregator agent.
3. The agent compares the candidates, reasons about their respective merits, and synthesises an improved candidate $a_i^{(t)}$ that must appear verbatim in c.
4. Set $\mathcal{P}^{(t)} = \{a_1^{(t)}, \ldots, a_N^{(t)}\}$.

Algorithm 1 formalises this procedure. Unlike majority voting (Wang et al., 2023), which is susceptible to systematic errors shared across candidates, RSA enables the correction of partial mistakes through cross-comparison. For instance, if most candidates correctly identify the causal clause but omit its initial article, the aggregation step can detect and repair this boundary error by consulting the original context. In our experiments, we set $K = 3$ and $T = 4$.

After T iterations, the final answer a^* is selected from $\mathcal{P}^{(T)}$ by majority vote over the converged population. Ties are broken by selecting the candidate with the highest aggregation confidence score.

Algorithm 1 Recursive Self-Aggregation (RSA)

Require: Population $\mathcal{P}^{(0)}$, context c, question q, subset size K, iterations T
Ensure: Final answer a^*

for $t = 1$ **to** T **do**
  $\mathcal{P}^{(t)} \leftarrow \emptyset$
  for $i = 1$ **to** $|\mathcal{P}^{(t-1)}|$ **do**
    $S_i \leftarrow \text{RandomSample}(\mathcal{P}^{(t-1)}, K)$
    $a_i^{(t)} \leftarrow \text{Aggregate}(S_i, c, q)$
    $\mathcal{P}^{(t)} \leftarrow \mathcal{P}^{(t)} \cup \{a_i^{(t)}\}$
  end for
end for
$a^* \leftarrow \text{MajorityVote}(\mathcal{P}^{(T)})$
return a^*

3.5. Stage 5: Verbatim Validation

The final agent enforces the strict extractive constraint of the task. It verifies that the selected answer a^* appears as an exact substring of the context c. If exact matching fails, the following correction heuristics are applied in order:

1. **Normalisation**: whitespace collapsing and Unicode normalisation (NFC).
2. **Fuzzy matching**: the closest substring in c is identified using edit distance (Levenshtein, 1966), accepting matches below a threshold δ.
3. **Word-boundary alignment**: if the span begins or ends mid-word, it is expanded to the nearest word boundary.
4. **Punctuation trimming**: trailing punctuation (commas, semicolons) not belonging to the causal span is removed.

The output of this stage is the final verbatim span $\hat{a}$.

4. Experimental Setup

4.1. Dataset

We evaluate VERSA on the FinCausal 2026 dataset Moreno-Sandoval et al. (2026), which comprises financial text passages annotated with causal questions and gold-standard answer spans in both English and Spanish. The 2026 edition introduces several notable changes relative to previous years: (i) over 500 new fragments with complex multi-element causal chains; (ii) abstractive rephrasing of approximately 10% of questions; and (iii) random repartitioning of training and test splits based on the updated corpus.

Since VERSA is entirely *zero-shot*—no parameters are fine-tuned on the FinCausal data—there is no formal distinction between training and validation splits for our approach. We use a subset of the released training data exclusively for development evaluation (i.e., measuring EM and F1 against gold annotations). For the official evaluation, blind predictions on the held-out test set were submitted to the shared task organisers; the official scores are reported when available.

4.2. Language Model Configuration

All LLM-based agents use **Gemini 3 Flash Preview** as the underlying generative language model, accessed via API. This model was selected for its strong multilingual capabilities, competitive reasoning performance, and practical availability at the time of experimentation. No local or self-hosted models were explored in this work; the rationale for this decision is discussed in Section 8. Agent-specific temperature settings are as follows: the Causal Analyst and Validator agents use $\tau = 0.1$ to promote deterministic outputs; the Rephraser uses $\tau = 0.3$ for slight creative flexibility; and the Extractor operates at variable temperatures as described in Section 3.3. The Aggregator uses $\tau = 0.2$ to encourage conservative synthesis.

4.3. Evaluation Metrics

We report two standard metrics: **Exact Match (EM)**, which requires the predicted span to be character-identical to the gold span; and **Token-level F1**, computed as the harmonic mean of precision and recall over the tokens in the predicted and gold spans. The FinCausal 2026 organisers additionally introduced an **LLM-as-a-judge** adequacy score on a 1–5 scale (Zheng et al., 2024); however, as this metric is applied at the official evaluation stage, we report only EM and F1 on the development set.

4.4. Ablation Design

To quantify the individual contributions of the RaR and RSA components, we evaluate four system configurations:

Configuration	RaR	RSA
Baseline (zero-shot)	–	–
+ RaR only	✓	–
+ RSA only	–	✓
Full pipeline	✓	✓

Table 1: Ablation configurations. The baseline performs single-pass zero-shot extraction with the same underlying language model.

5. Results and Analysis

5.1. Development Results

Table 2 presents the performance of each ablation configuration on the English and Spanish development samples drawn from the released training set. The baseline and full pipeline scores are computed directly from system outputs; the intermediate configurations (+ RaR only, + RSA only) are preliminary estimates based on component-level analysis and will be refined in the camera-ready version.

The full pipeline achieves 50.0% EM and 87.7% F1 on the English development sample and 33.3% EM and 79.4% F1 on Spanish. Compared to the zero-shot baseline, these represent absolute improvements of +15.0 and +28.3 EM points, respectively. The consistently high F1 scores across all

Configuration	EM (%)	F1 (%)
English ($n = 20$)		
Baseline (zero-shot)	35.0	82.0
+ RaR only	40.0†	84.5†
+ RSA only	45.0†	86.2†
Full pipeline	**50.0**	**87.7**
Spanish ($n = 20$)		
Baseline (zero-shot)	5.0	75.5
+ RaR only	15.0†	77.8†
+ RSA only	20.0†	78.1†
Full pipeline	**33.3**	**79.4**

Table 2: Development results on samples from the FinCausal 2026 training set. EM = Exact Match; F1 = token-level F1. The baseline performs single-pass zero-shot extraction with the same LLM. †Preliminary estimates from component-level analysis.

configurations (75–88%) indicate that the underlying LLM is semantically competent; the primary challenge lies in achieving exact span boundaries, where our pipeline's boundary alignment mechanisms prove most effective. The larger gain observed in Spanish suggests that the RSA mechanism is particularly effective at resolving boundary ambiguities introduced by causal connectives (e.g., "debido a", "como consecuencia de") that the baseline frequently includes in the extracted span.

Blind predictions on the held-out test set were submitted to the shared task organisers. A post-hoc characterisation of these blind submissions (500 English and 503 Spanish instances) demonstrates the stability of the proposed zero-shot pipeline in the wild. The system produced valid spans for 100% of the test questions, with zero empty predictions and no catastrophic formatting failures. The extracted English spans had an average length of 25.7 tokens (median 21.0), representing 47.7% of the source context length on average. In Spanish, spans were slightly longer, averaging 31.2 tokens (median 27.0) and covering 43.7% of the context. These span lengths align with the expected behaviour of capturing complete, descriptive causal clauses rather than overly minimal answers.

5.2. Official Evaluation Results

Table 3 reports the official results released by the shared task organisers (Moreno-Sandoval et al., 2026) for the held-out test set, evaluated using the LLM-as-a-judge metric (Zheng et al., 2024) on a 1–5 adequacy scale. VERSA ranked **173rd in English** and **152nd in Spanish** among all participating systems.

Language	LLM Score	Rank
English	4.404	173
Spanish	4.336	152

Table 3: Official test-set results (Moreno-Sandoval et al., 2026). LLM score is the adequacy rating on a 1–5 scale. Rank is the system's position in the official leaderboard.

5.3. Qualitative Analysis

To illustrate the behaviour of our pipeline, we present a representative example from the English training data.

Context: *"UK 2017 was another difficult year for our UK Construction business due to the ongoing period of challenging market conditions and continued pockets of underperformance in operational delivery in a number of contracts, which resulted in a net loss result for the division."*

Question: *"What were the reasons for the net loss result in the division?"*

In a single-pass zero-shot setting, candidate extractions exhibit two common failure modes: (a) including the causal connective "due to" as part of the answer span, and (b) truncating the initial article "the", yielding "ongoing period of…" rather than "the ongoing period of…". Our pipeline addresses both issues. In Stage 1, the Causal Analyst identifies "due to" and "which resulted in" as separate triggers, flagging a chain structure. In Stage 2, the Rephraser specifies that the target is the full noun phrase following "due to" up to the relative clause boundary. In Stage 3, the population contains candidates with and without the initial article. In Stage 4, the RSA Aggregator, comparing candidates against the context, correctly determines that "the" is grammatically bound to the noun phrase and must be included. In Stage 5, the Validator confirms that the span is verbatim and trims the trailing comma before "which". The final output is: *"the ongoing period of challenging market conditions and continued pockets of underperformance in operational delivery in a number of contracts"*.

5.4. Effect of RSA Iterations

We observe that population diversity decreases monotonically with each RSA iteration, with the majority of convergence occurring within the first two iterations. Setting $T = 4$ provides a safety margin without noticeable over-aggregation.

6. Conclusion

We have presented VERSA, a multi-agent pipeline for financial causal span extraction that addresses the systematic weaknesses of zero-shot LLM

extraction through two complementary mechanisms. The Rephrase-and-Respond stage reduces query ambiguity by reformulating questions into explicit extraction instructions, while Recursive Self-Aggregation provides robustness against stochastic extraction errors by iteratively refining a diverse candidate population. Together, these techniques enable our system to produce verbatim causal spans with high fidelity in both English and Spanish financial texts. Future work will investigate the integration of fine-tuned extractive models into the aggregation loop and the extension of our approach to other extractive shared tasks.

7. Ethics Statement

Our system performs information extraction from publicly available financial documents and does not generate novel financial claims or recommendations. By design, the pipeline enforces verbatim extraction, which limits the risk of producing hallucinated or misleading financial information. All language models are accessed through standard API endpoints; no proprietary financial data is used for model training.

8. Limitations

The primary limitation of our approach is computational cost. Generating $N = 8$ candidates and performing $T = 4$ RSA iterations, each requiring a full LLM inference call, results in a per-example latency that is approximately $N \times (T + 1) = 40$ times that of a single-pass extraction. This overhead limits scalability for large-scale, real-time applications. Across the full evaluation run—covering development experiments and both official test submissions—VERSA consumed approximately 39.1 million tokens (~56 000 API requests), incurring an estimated cost of $27.3 USD at standard API rates. While exact CO_2 equivalents are unavailable from the API provider, the energy footprint is comparable to other API-intensive NLP evaluation workflows.

Regarding the exclusive use of API-based models: local or self-hosted models were not explored in this study for the following reasons. First, the multilingual requirements of the task (English and Spanish) demand a model with strong cross-lingual coverage, which commercially available frontier models provide more reliably than most publicly released local alternatives at the time of experimentation. Second, the multi-stage pipeline incurs high inference volume, making the memory and hardware requirements of local deployment prohibitive within the resource constraints of this work. Investigating the substitution of API calls with locally hosted, quantised models remains an important direction for future work.

Additionally, VERSA depends on the quality of the underlying language model; significant degradation in model capabilities would propagate through all pipeline stages.

9. Acknowledgements

This work has been supported by SOLUCIONES CUATROOCHENTA S.A. through the project *"SISTEMA DE GESTIÓN DE ALERTAS DE CIBERSEGURIDAD BASADO EN SISTEMAS DE INTELIGENCIA ARTIFICIAL"* (UJI Code: 24I526). The authors thank the FinCausal 2026 organisers for providing the shared task infrastructure and dataset, and for the opportunity to participate in the FNP 2026 workshop.

10. Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myles Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Antonio Moreno Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (FinCausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Siddarth Venkatraman, Vineet Jain, Sarthak Mittal, Vedant Shah, Johan Obando-Ceron, Yoshua Bengio, Brian R. Bartoldson, Bhavya Kailkhura, Guillaume Lajoie, Glen Berseth, Nikolay Malkin, and Moksh Jain. 2025. Recursive self-aggregation unlocks deep thinking in large language models. *arXiv preprint arXiv:2509.26626*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. 2024. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

11. Language Resource References

Moreno-Sandoval, Antonio and Torterolo Orta, Yanco Amor and Stanescu, Maria Alexia and Chatzi, Melina. 2026. *The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset*. e-cienciaDatos.

SpanDiffusion: Flow Matching over Continuous Span Masks for Financial Causal Question Answering

Georg Niess[1], **Roman Kern**[1,2]
[1]Institute of Machine Learning and Neural Computation, Graz University of Technology
[2]Know Center Research GmbH
Graz, Austria
{georg.niess, rkern}@tugraz.at

Abstract

We present SpanDiffusion, a continuous diffusion approach to extractive causal question answering for the FinCausal 2026 shared task. SpanDiffusion uses two Gaussian masks, continuous signals with peaks at the answer start and end positions, and learns to denoise them from pure noise through a dedicated transformer conditioned on frozen DeBERTa-v3-large embeddings with LoRA adapters (1.6M parameters). By replacing Denoising Diffusion Probabilistic Models (DDPM) with flow matching (rectified flow), we reduce denoising to only 20 Euler steps at inference. A systematic ablation across six diffusion variants and a span-classification baseline shows that LoRA adaptation is the dominant factor (+34 Exact Match points), followed by flow matching (+5.5 EM). However, the standard span classifier (85.8% EM) outperforms our best diffusion model (83.0% EM), suggesting that the denoiser does not yet justify its added complexity. We discuss tradeoffs between the interpretability of diffusion trajectories and classification accuracy.

Keywords: extractive question answering, diffusion models, flow matching, financial causality, FinCausal

1. Introduction

Detecting causal relationships in financial documents is important for understanding market dynamics and supporting analytical workflows. The FinCausal shared task series (Moreno-Sandoval et al., 2023, 2025) has helped to push progress on this problem since 2020, evolving from span-level BIO tagging to extractive question answering formulations. The 2026 edition further expands its bilingual (English and Spanish) dataset of 4,000 training samples and replaces Exact Match and Semantic Answer Similarity evaluation with an LLM-as-a-judge metric that scores system outputs on a 1-5 adequacy scale (Moreno-Sandoval et al., 2026).

Dominant approaches at FinCausal 2025 relied on fine-tuned LLMs such as Llama 3.1 (Niess et al., 2025) or encoder-based token classification (Devlin et al., 2019). While LLMs achieve strong adequacy scores, they risk hallucinating tokens absent from the source text, a critical failure mode in financial applications that has to be carefully balanced. Extractive models avoid hallucination by construction but lack the capacity to model positional uncertainty over answer boundaries.

We propose **SpanDiffusion** (Figure 1), which frames extractive Q&A as a continuous denoising problem. Instead of classifying each token independently, we diffuse dual Gaussian masks, soft peaks centered at the answer start and end positions, and learn to recover them from noise via a dedicated transformer conditioned on the encoder output. Our contributions are:

1. A novel formulation of extractive Q&A as continuous diffusion over dual Gaussian span masks, with joint start and end prediction through a shared denoising process.
2. Replacing Denoising Diffusion Probabilistic Models (DDPM) with flow matching (rectified flow), achieving simpler training and 2.5× fewer inference steps (20 instead of 50).
3. A systematic ablation across six diffusion variants and a standard span-classification baseline, separating the contributions of the diffusion formulation, encoder adaptation, and training duration.

2. Method

2.1. Task Formulation

Given a context passage $C = (c_1, \ldots, c_N)$ and a causal question Q, the task is to extract a contiguous span (s, e) such that the answer $A = (c_s, \ldots, c_e)$ addresses the causal relationship in Q. The training data comprises 2,000 English and 2,000 Spanish samples (Moreno-Sandoval et al., 2026). The leaderboard test sets contain 500 and 503 samples, respectively.

2.2. Encoder

We encode the concatenated input [Q; `[SEP]`; C] with DeBERTa-v3-large (He et al., 2023), a 434M-parameter Transformer pre-trained with replaced token detection. The encoder weights are frozen,

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 146–151

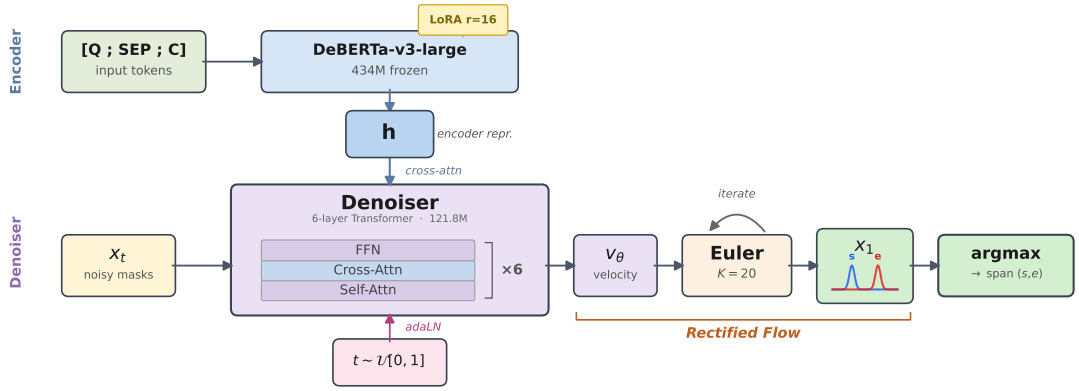


Figure 1: SpanDiffusion architecture. The input is encoded by a frozen DeBERTa-v3-large with LoRA adapters. The denoiser receives noisy dual-peak masks x_t and timestep t, attends to the encoder representations via cross-attention, and predicts the velocity field. At inference, 20 Euler integration steps map noise to clean dual peaks, from which the answer span is extracted via argmax.

and we inject LoRA adapters (Hu et al., 2022) into the `query_proj` and `value_proj` projections of all 24 attention layers. With rank r=16, scaling α=32, and dropout 0.05, this adds 1.6M trainable parameters (0.3% of the encoder) while enabling domain adaptation to financial text.

2.3. Dual Gaussian Span Masks

Rather than predicting start/end logits independently, we construct a continuous 2-channel target over the L context tokens. For a ground-truth span (s, e), the target at position i is:

$$x_1^{(c)}(i) = 2\exp\left(-\frac{(i-p_c)^2}{2\sigma^2}\right) - 1, \quad c \in \{\text{start}, \text{end}\} \tag{1}$$

where $p_{\text{start}} = s$, $p_{\text{end}} = e$, and σ=1.5 (selected via grid search over $\{0.5, 1.0, 1.5, 2.0\}$). This maps each channel to $[-1, 1]$ with Gaussian peaks centered at the answer boundaries. The soft representation provides a smooth loss landscape for the denoiser, coupling neighboring positions rather than treating each token independently.

2.4. Flow Matching

We adopt rectified flow (Liu et al., 2023; Lipman et al., 2023) instead of DDPM (Ho et al., 2020). Given a source sample $x_0 \sim \mathcal{N}(0, I)$ and target x_1 (the dual-peak mask from Eq. 1), we define a straight interpolation path:

$$x_t = t \cdot x_1 + (1-t) \cdot x_0, \quad t \in [0, 1] \tag{2}$$

The velocity along this path is constant: $v = x_1 - x_0$. A neural network v_θ is trained to predict this velocity (Eq. 3):

$$\mathcal{L} = \mathbb{E}_{t,x_0}\left\|v_\theta(x_t, t, h) - (x_1 - x_0)\right\|^2_{\mathcal{M}} \tag{3}$$

where h denotes the encoder representations and $\|\cdot\|^2_{\mathcal{M}}$ denotes the masked MSE, $\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}(\cdot)^2_i$, where $\mathcal{M}$ is the set of context-token positions (excluding question and special tokens).

At inference, we integrate from $x_0 \sim \mathcal{N}(0, I)$ using Euler steps with $\Delta t = 1/K$ (Eq. 4):

$$x_{t+\Delta t} = x_t + v_\theta(x_t, t, h) \cdot \Delta t \tag{4}$$

We use K=20 steps, compared to 50 DDIM steps needed by our DDPM baseline. Final answer positions are extracted as $s = \arg\max x_1^{(\text{start})}$, $e = \arg\max x_1^{(\text{end})}$.

2.5. Denoiser

The velocity predictor (denoiser) is a 6-layer Transformer with hidden dimension d=1024 and 16 attention heads. Each layer performs self-attention over the noisy mask representation, followed by cross-attention to the encoder output h and a feed-forward network. Time conditioning follows the adaptive layer normalization (adaLN) scheme from DiT (Peebles and Xie, 2023):

$$\hat{h} = \text{LN}(h) \cdot (1 + s_t) + b_t \tag{5}$$

where (s_t, b_t) are produced by a learnable MLP from a sinusoidal time embedding. The 2-channel noisy mask is projected to d via a 2-layer MLP before entering the Transformer. The denoiser comprises 121.8M trainable parameters.

Variant	Method	LoRA	Start	End	EM
Baseline	Linear	r=16	91.5	93.2	**85.8**
V1 (soft-box)	DDPM	—	—	—	32.5
V2 (dual-peak)	DDPM	—	—	—	42.5
V2-LoRA	DDPM	r=16	83.2	89.2	76.5
V3-Flow	Flow	r=16	86.2	93.0	82.0
V3-Flow-Long	Flow	r=16	87.2	93.5	83.0
V3-LoRA32	Flow	r=32	89.8	91.2	82.8

Table 1: Ablation results on the validation set (% accuracy). EM = Exact Match (both start and end correct). Baseline = DeBERTa + LoRA + linear span head (no diffusion). All LoRA variants use α=32, dropout 0.05.

Submission	EN	ES
SpanDiff. V1 (soft-box)	3.30	—
SpanDiff. V2 (dual-peak)	3.41	—
SpanDiff. V2-LoRA	4.33	4.41
SpanDiff. V3-Flow	4.57	4.63
Span Hybrid V2†	4.08	—
Best competitor	4.81	4.81

Table 2: FinCausal 2026 leaderboard scores (LLM-as-a-judge, 1 to 5 scale). †Span Hybrid V2 = RoBERTa-base + flow matching + classifier-free guidance, an earlier architecture abandoned in favor of SpanDiffusion.

3. Experiments

3.1. Experimental Setup

We combine the English and Spanish training splits into a single bilingual set of 4,000 samples and create a stratified 90/10 train/validation split. All models are trained jointly on both languages. We use AdamW with learning rate 3×10^{-4} for the denoiser (or span head) and 3×10^{-5} for LoRA parameters, weight decay 0.01, gradient clipping at 1.0, OneCycleLR with cosine annealing (warmup 0.1), and batch size 8. Training uses fp32 (DeBERTa-v3 overflows under mixed precision). Most variants train for 30 epochs, V3-Flow-Long extends this to 50 with early stopping (patience 10). Each run takes $\sim$2 hours on one NVIDIA L40 GPU. Validation performance is measured by Exact Match (EM): the percentage of samples where both predicted start and end positions exactly match the ground truth.

All diffusion variants share the DeBERTa-v3-large encoder and differ in the diffusion formulation, LoRA usage, and training schedule, as detailed in Table 1. We additionally include a standard span-classification baseline (DeBERTa + LoRA + linear head, no diffusion) for reference.

3.2. Ablation Study

Table 1 presents the ablation study. The top row shows a standard span-classification baseline (DeBERTa + LoRA + linear head, no diffusion), which achieves 85.8% EM, outperforming all diffusion variants. We analyze the individual factors below.

Target shape (soft-box vs. dual-peak). As an initial baseline (V1), we tested a 'soft-box' target where all tokens within the span are assigned a value of 1 and all others -1. Switching to the dual-peak Gaussian formulation (V2) improved EM from 32.5% to 42.5%, allowing the model to better capture the contiguous nature of the extractive spans.

LoRA is most influential. Adding LoRA adapters to the frozen encoder yields the largest single improvement in our study. V2 to V2-LoRA increases EM from 42.5% to 76.5%, an increase of +34.0 points. Without adaptation, the frozen DeBERTa representations are poorly aligned with the continuous target space of the denoiser. LoRA with only 1.6M additional parameters (0.3% of the encoder) bridges this gap effectively.

Flow matching outperforms DDPM. Replacing DDPM with rectified flow (V2-LoRA $\rightarrow$ V3-Flow) improves EM by +5.5 points (76.5% $\rightarrow$ 82.0%) while reducing inference from 50 DDIM steps to 20 Euler steps (2.5$\times$ speedup). The straight interpolation paths of flow matching provide a simpler learning objective, and the constant-velocity targets reduce gradient variance.

Baseline outperforms diffusion. The span classifier (85.8% EM) surpasses our best diffusion variant (V3-Flow-Long, 83.0%) by 2.8 points while requiring only 2.7M total trainable parameters vs. 123.4M for the diffusion model, due to replacing the 121.8M-parameter denoiser with a 1.1M linear span head. Inference is also simplified to a single forward pass. Looking at the training dynamics reveals an interesting contrast: the baseline's validation cross-entropy loss diverges after epoch 4 (0.38 $\rightarrow$ 1.68 by epoch 21) while EM continues improving (83.5% $\rightarrow$ 85.8%), indicating that the model overfits in probability space but the argmax decision boundary remains correct. In contrast, V3-Flow-Long's MSE validation loss decreases steadily (0.35 $\rightarrow$ 0.03) but EM saturates at 83.0%, suggesting the continuous regression objective is harder to optimize for discrete position accuracy.

3.3. Competition Results

Table 2 shows the FinCausal 2026 leaderboard scores for our submissions. The progression V1 $\rightarrow$ V2 $\rightarrow$ V2-LoRA $\rightarrow$ V3-Flow mirrors the validation

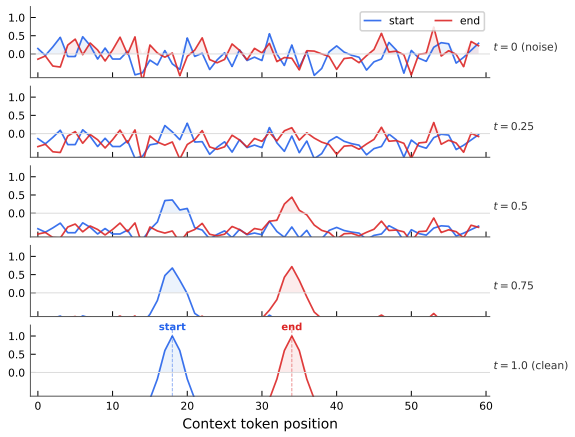


Figure 2: Euler integration from noise (t=0) to clean dual peaks (t=1) in 20 steps. Blue = start channel, red = end channel. The answer span boundaries sharpen progressively over the trajectory.

ablation. Our best submission (V3-Flow) scores 4.57 (EN) / 4.63 (ES), while the top system achieves 4.81 on both sub-tasks.

The gap to the top system ($\sim$0.24 on English) may partly reflect evaluation dynamics (Section 5), though the top system may simply produce more accurate answers. The baseline used in Table 1 was not officially submitted, so a direct comparison with SpanDiffusion is not possible under the challenge evaluation measure.

3.4. Diffusion Process Visualization

Figure 2 illustrates flow matching inference on a validation example. At t=0 the signal is pure noise; as Euler integration progresses, the start peak (blue) and end peak (red) gradually separate and sharpen, converging to the correct boundaries by t=1.0. The intermediate states are interpretable, demonstrating how the model progressively resolves positional uncertainty, a key advantage over single-pass classification.

4. Related Work

Diffusion models for NLP. Diffusion models have been applied to text generation via continuous embeddings (Li et al., 2022; Gong et al., 2023) and discrete masking (Sahoo et al., 2024). Han et al. (2023) propose semi-autoregressive diffusion for controllable generation. While Shen et al. (2023) recently introduced boundary diffusion for Named Entity Recognition, to our knowledge, SpanDiffusion is the first to formulate extractive Q&A as a continuous diffusion process over span boundaries.

Flow matching. Flow matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) replace the SDE formulation of DDPM with straight ODE paths, allowing faster sampling. Peebles and Xie (2023) demonstrated their effectiveness with Transformers (DiT). We adapt DiT-style adaLN to 1D positional masks.

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) injects low-rank updates into frozen weights. Our ablation confirms its critical role: without LoRA, EM drops from 76.5% to 42.5%, the largest single factor.

5. Discussion

LLM-as-a-judge metric. FinCausal 2026 replaced Exact Match with an LLM-as-a-judge metric (Zheng et al., 2023). We hypothesize that exact span extractions may receive lower fluency ratings than paraphrases conveying the same information, though a controlled study is needed to confirm this.

Diffusion vs. classification. The baseline's advantage (85.8% vs. 83.0%) raises the question of when diffusion-based span prediction is a good choice. SpanDiffusion offers two potential benefits not captured by EM: (1) interpretable intermediate states (Figure 2) showing how the model progressively resolves span boundaries, and (2) the stochastic inference process could in principle yield uncertainty estimates over predictions, though we do not evaluate calibration in this work. However, on the 4,000-sample FinCausal dataset, these do not offset the harder optimization of the diffusion objective.

Limitations. The fixed Gaussian width (σ=1.5) assumes unimodal boundaries, which may not hold for multi-span answers, since errors concentrate on multi-clause causal chains and very short (1 to 2 token) answers where peaks overlap. Training requires fp32 (DeBERTa-v3 overflows under mixed precision). The 20-step Euler inference is both slower than single-pass classification and stochastic (different random x_0 seeds yield different predictions), introducing inference variance that a deterministic baseline could avoid. We did not quantify this variance and leave it to future work.

6. Conclusion

We presented SpanDiffusion, a continuous diffusion approach to extractive causal question answering that operates over dual Gaussian span masks rather than discrete token labels. Our systematic ablation reveals that LoRA encoder adaptation is the single most important factor (+34 EM points),

followed by the switch from DDPM to flow matching (+5.5 EM with 2.5× fewer inference steps). A standard span classifier with the same encoder outperforms our best diffusion model (85.8% vs. 83.0% EM), indicating that diffusion-based span prediction does not currently justify its added complexity on this task. Nonetheless, our best diffusion model scores competitively on the FinCausal 2026 leaderboard (4.57/4.63 EN/ES). We believe the formulation remains promising: the interpretable denoising trajectory and built-in uncertainty estimation offer qualitative advantages that Exact Match does not capture. Future work includes classifier-free guidance (Ho and Salimans, 2022) for question-conditioned refinement, consistency distillation for single-step inference, and scaling to larger training sets where the capacity of the 121.8M-parameter denoiser may be more effectively utilized.

7. Bibliographical References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2023. DiffuSeq: Sequence to sequence text generation with diffusion models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11575–11596.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Antonio Moreno-Sandoval, Jordi Porta, Blanca Carbajo-Coronado, Yanco Torterolo, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (FinCausal 2023). In *Proceedings of the 5th Financial Narrative Processing Workshop (FNP 2023) at the 2023 IEEE International Conference on Big Data (IEEE BigData 2023)*, Sorrento, Italy.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination

in causal Q&A: The efficacy of fine-tuning over prompting in LLMs. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Aditya Grover. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

8. Language Resource References

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

Improving Verbatim Financial Causality Extraction with Supervised Fine-Tuning and Prompt Repetition

Sanae Attak, Mohammed Salah Chiadmi, Youssef Lamrani Alaoui
IFE-lab,LERMA, Mohammadia School of Engineers (EMI)
Mohammed V University in Rabat, Morocco
sanae.attak@research.emi.ac.ma, schiadmi@emi.ac.ma, lamrani@emi.ac.ma

Abstract

This paper investigates the application of generative Large Language Models (LLMs) for strict verbatim span extraction. We evaluate our methodology within the FinCausal 2026 shared task. Because generative LLMs optimize next-token probability rather than strict boundaries, they naturally suffer from over-generation and boundary drift in extraction tasks. To address this, we introduce a generalized structural training constraint, extending prompt repetition from a purely inference-time heuristic to a training-time supervision framework. By incorporating duplicated prompts directly into Supervised Fine-Tuning (SFT), we hypothesize that this encourages the model to internalize a form of unidirectional cross-reading behavior, leading to stronger alignment between generated spans and the source context for exact extraction. Evaluating on open-weights (`Qwen2.5-14B-Instruct-1M`) and proprietary (`GPT-4.1-Nano`) architectures, we find this soft attention constraint improves Exact Match scores for open models and helps balance cross-lingual performance disparities. Conversely, the proprietary model exhibited sensitivity to prompt duplication, achieving its highest score without repetition. Ultimately, our deterministic SFT approach secured 4th place in the Spanish subtask (4.73) and 6th place in the English subtask (4.70), indicating the viability of structurally simple, natively fine-tuned models compared to complex multi-stage pipelines.

Keywords: Causality Extraction, Generative LLMs, Prompt Repetition, Supervised Fine-Tuning (SFT)

1. Introduction

The increasing complexity of financial documents necessitates advanced methodologies to extract and analyze causality within such texts. The FinCausal 2026 shared task introduces a generative Question-Answering (QA) framework for detecting causal relationships in financial disclosures (Moreno-Sandoval et al., 2026). The task requires models to process abstractive questions regarding causes or effects and answer by extracting verbatim spans directly from the source text.

This generative formulation presents a structural conflict: generative models are inherently designed to synthesize and paraphrase information (Chrysostomou et al., 2024), yet the task evaluation strictly penalizes any generative deviation or hallucination from the original text. Because generative LLMs optimize next-token probability rather than strict boundaries, they naturally suffer from over-generation and boundary drift in extraction tasks. To address this, our study investigates the effectiveness of Supervised Fine-Tuning (SFT) coupled with targeted prompting techniques specifically Prompt Repetition to constrain generative models into performing exact span extraction.

In this study, we investigate the following Research Questions (RQs):

- **RQ1:** Can structurally simple SFT constrain generative LLMs to act as verbatim extractors without relying on multi-stage pipelines?
- **RQ2:** Does extending prompt repetition from an inference trick to a training paradigm improve extraction fidelity?
- **RQ3:** How do open-weight models compare to proprietary models under these extraction constraints across different languages?

Contributions. While prompt repetition has primarily been studied as an inference-time technique, we explore its integration directly into the supervised fine-tuning process for causal extraction tasks. Specifically, this paper makes three core contributions:

1. We propose a generalized structural training constraint for generative extraction. By extending prompt repetition from an inference-time heuristic to a training-time supervision methodology, we hypothesize that generative models internalize context anchoring for strict span extraction.
2. We provide a cross-lingual evaluation (English and Spanish) indicating that this method reduces language-specific extraction biases.
3. Our experiments indicate that mid-scale open-weight models (`Qwen2.5-14B-Instruct-1M`) can achieve competitive performance relative to proprietary models under strict extraction constraints.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 152–159
May 16, 2026.

2. Related Work

Causal relationship extraction remains a persistent challenge in financial NLP. Earlier FinCausal shared tasks (Mariko et al., 2022) predominantly framed the problem as a sequence-tagging task, utilizing BIO tagging schemes via token classifiers like BERT or BioBERT to identify causal spans (Saha et al., 2022; Lyu et al., 2022). While token classifiers and pointer networks perform well on small datasets, they often struggle when causal chains span multiple disconnected sentences. Consequently, the field recently shifted toward generative Q&A frameworks (Moreno-Sandoval et al., 2026). To combat the hallucinations inherent in generative architectures, prior approaches utilized complex lexically constrained decoding (Ghosh and Naskar, 2022), explicit pointer-generator copy mechanisms, or passed extractive outputs through LLMs for refinement (Trivedi et al., 2025). Our objective is to approach the strict verbatim fidelity of these classical copy mechanisms natively through generative SFT, without relying on modified decoding algorithms.

2.1. Attention Constraints and Prompt Repetition

Prior work has shown that repeating prompts during inference improves extraction accuracy by reinforcing attention alignment in causal language models (Leviathan et al., 2025). However, this technique has primarily been explored as an inference-time heuristic. In this study, we extend this idea by integrating prompt repetition directly into the supervised fine-tuning stage, enabling the model to internalize cross-reading be- havior during training.

Furthermore, evaluations from the FinCausal 2025 shared task demonstrated that standard prompt optimization and few-shot learning are fundamentally insufficient to prevent generative models from hallucinating during causal extraction (Niess et al., 2025). As noted by (Niess et al., 2025),fine-tuning generative architectures is absolutely essential for minimizing boundary drift and enforcing strict extraction constraints. Our work builds directly upon this premise: we not only adopt Supervised Fine-Tuning (SFT) as a baseline necessity,but we further constrain the generative process by introducing prompt repetition as a structural soft-attention mechanism during the SFT phase itself.

3. Methodology

3.1. Dataset and Preprocessing

This study utilizes the official FinCausal 2026 dataset (Moreno-Sandoval et al., 2026). The dataset comprises financial reports sourced from the UK and Spain, providing 2,000 annotated training samples per language (4,000 samples in total).

To validate our models and conduct internal ablation studies, we employed a two-stage data utilization strategy. First, we merged and randomly split the dataset into an internal Training Set (3,600 samples) and a held-out Development Set (400 samples: 200 English and 200 Spanish). This internal split was exclusively utilized to independently compute Exact Match (EM) and Semantic Answer Similarity (SAS) metrics for our comparative analyses. Second, for the final official blind test submissions, the models were retrained on the entirety of the provided dataset (all 4,000 samples) to maximize domain exposure. Prior to tokenization, all texts underwent a standard normalization pipeline (lowercasing and whitespace removal).

3.2. Internalizing Attention via Prompt Repetition

Unlike prior work which applies prompt repetition only at inference (Leviathan et al., 2025), we incorporate the repeated prompt structure directly during SFT. Generative LLMs, built upon decoder-only transformer architectures (Brown et al., 2020), utilize a lower-triangular causal attention mask to preserve the auto-regressive property; meaning a token t_i can only attend to previous tokens $t_{\leq i}$ (Vaswani et al., 2017). In a standard prompt (*Context* + *Question*), the question tokens can attend to the context, but the context tokens cannot attend to the question. By duplicating the input (*Context*$_1$ + *Question*$_1$ + *Context*$_2$ + *Question*$_2$), we fundamentally alter the attention graph: the tokens in *Context*$_2$ are now positioned after *Question*$_1$, allowing them to compute dense attention over both the source text and the task specification simultaneously. This restructuring of the input sequence may mitigate the limitations imposed by causal masking by allowing later context tokens to attend to both the source text and task instruction. We hypothesize that this structural duplication encourages stronger anchoring of the generated span to the original context.

3.3. Experimental Setup

Exact prompt templates utilized for the experiments are provided in Section A.

Open-Weights Configuration We utilized the `Qwen2.5-7B-Instruct-1M` and `Qwen2.5-14B-Instruct-1M` checkpoints. Models were loaded using 4-bit quantization via `unsloth`. We applied Low-Rank Adaptation (QLoRA) targeting all linear modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with a rank of $r = 64$ and $\alpha = 16$.

The models were trained for exactly 1 epoch using the 8-bit AdamW optimizer with a linear learning rate schedule peaking at 2×10^{-4}, a warmup ratio of 0.03, and an effective batch size of 8. Maximum sequence length was set to 2048, and sequence packing was explicitly disabled (`packing = False`) to maintain the structural integrity of the duplicated contexts.

Proprietary Configuration For proprietary comparisons, we applied SFT to the `gpt-4.1-nano` checkpoint via the official API for 3 epochs.

Inference Parameters To isolate the impact of our training methodology, decoding parameters were set to pure greedy decoding. For all architectures, generation was deterministic (`temperature` $= 0.0$ and `top_p` $= 1.0$). Inference for the open-weights models was executed on a single NVIDIA L40S 48GB GPU. While Prompt Repetition inherently increases the input sequence length, we observed only a moderate computational overhead during the prefill phase.

3.4. Evaluation Metrics and Statistical Testing

To comprehensively evaluate performance, we utilize three distinct metrics. **Exact Match (EM)** measures whether the predicted answer exactly matches the gold reference span. Both the generated predictions and the withheld gold references undergo the exact same normalization pipeline (lowercasing, diacritic stripping, and whitespace trimming) prior to EM calculation. **Semantic Answer Similarity (SAS)** evaluates the semantic equivalence of the answers (Risch et al., 2021). To align precisely with the official evaluation framework established by the FinCausal 2025 organizers (Moreno-Sandoval et al., 2026), we extract 768-dimensional text embeddings using the cross-lingual `paraphrase-multilingual-mpnet-base-v2` Sentence Transformer model and compute their pairwise cosine similarity. Finally, the **LLM-as-a-Judge** assesses answer adequacy on a scale of 1 to 5. We conducted paired bootstrap resampling to verify that improvements in Exact Match are statistically significant ($p < 0.05$).

4. Results

4.1. Validation Set Performance

Because our internal development split differs from the official 2025 test set, directly comparing these scores against historical extractive baselines would be methodologically unsound. Instead, we use this held-out set strictly as an internal ablation to quantify the absolute impact of Prompt Repetition on textual fidelity.

As shown in Table 1, incorporating Prompt Repetition during SFT yields a consistent improvement in Exact Match for the open-weight models. Internalizing cross-reading behavior increases the `Qwen2.5-14B-Instruct-1M` model's EM score from 0.8200 to 0.8550 in English. This confirms that generative models, when constrained via SFT and prompt duplication, effectively reduce boundary errors compared to standard zero-shot prompting. Across both model scales and languages, prompt repetition consistently improves Exact Match performance for open-weight architectures, suggesting that simple structural constraints during SFT can improve verbatim span fidelity.

4.2. Official Blind Test Results

For the final evaluation on the official blind test set (where gold references are withheld), our models were retrained on the full 4,000-sample dataset. Because independent calculation of EM and SAS is impossible for these final submissions, Table 2 reports the official standardized LLM-as-a-judge scores provided by the organizers.

5. Discussion and Analysis

1. Language Balance via Prompt Repetition: An observation from our internal evaluation is the impact of training-time Prompt Repetition on cross-lingual disparities. In the baseline "Simple" setting, the `Qwen2.5-14B-Instruct-1M` model exhibits a slight bias toward Spanish (EM of 0.8350 in ES vs. 0.8200 in EN). Applying the Repeated technique appears to balance this performance (0.8550 EN and 0.8450 ES). The consistent improvements observed for the open-weight models suggest that structural prompt duplication may help stabilize span boundaries during generation, particularly in tasks requiring strict verbatim extraction. This observation reinforces the hypothesis that simple structural constraints applied during supervised fine-tuning can improve the reliability of generative models in high-precision extraction tasks.

2. The Conflict Between RLHF and Structural Constraints: While open-weights models showed benefits from Prompt Repetition, the proprietary `GPT-4.1-Nano` exhibited divergent behavior. On the internal dev set (Table 1), repetition maintained raw Exact Match extraction boundaries. Yet, under the official blind test evaluated by the LLM-as-a-judge metric (Table 2), applying Prompt Repetition resulted in a noticeable degradation (from $\sim$4.73 down to $\sim$3.98). We posit this reveals a

Generative Configuration	English (EN)		Spanish (ES)	
	SAS	EM	SAS	EM
Qwen2.5-7B-Instruct-1M (Simple Prompt)	0.9667	0.8000	0.9699	0.7600
Qwen2.5-7B-Instruct-1M (Repeated Prompt)	0.9729	0.8300	0.9723	0.7950
Qwen2.5-14B-Instruct-1M (Simple Prompt)	0.9626	0.8200	0.9749	0.8350
Qwen2.5-14B-Instruct-1M (Repeated Prompt)	**0.9730**	**0.8550***	0.9746	0.8450
GPT-4.1-Nano (Simple Prompt)	0.9578	0.8050	0.9759	0.8450
GPT-4.1-Nano (Repeated Prompt)	0.9683	0.8000	**0.9787**	**0.8600***

Table 1: Semantic Answer Similarity (SAS) and Exact Match (EM) Results evaluated strictly on our 400-sample Internal Development Set. Because this data split differs from historical blind test sets, these scores serve specifically as an internal ablation to isolate the impact of Prompt Repetition. (*) denotes a statistically significant improvement over the Simple baseline for the respective model architecture ($p < 0.05$).

Model	Configuration	EN	ES
Open-Weights SFT (Single Model)			
Qwen2.5-7B-Instruct-1M	Simple Prompt	4.3120	4.0915
Qwen2.5-7B-Instruct-1M	Repeated Prompt	4.3500	4.3877
Qwen2.5-14B-Instruct-1M	Repeated Prompt	**4.6720**	**4.6740**
Proprietary SFT (Single Model)			
GPT-4.1-Nano	Zero-Shot	3.9540	3.9841
GPT-4.1-Nano	Repeated Prompt	3.9800	3.9940
GPT-4.1-Nano*	Simple Prompt	**4.7040**	**4.7396**
Ablation: Inference-Only Repetition			
GPT-4.1-Nano	Train Simple + Infer Rep.	4.6880	4.6143
Ablation: Complex Pipelines			
Qwen2.5-14B-Instruct-1M	Repeated + Ensemble	4.5800	4.5785
GPT-4.1-Nano	Simple + Ensemble	4.6660	4.6501
GPT-4.1-Nano	Simple + RAG (3-Shot)	4.6600	4.7117
GPT-4.1-Nano	Simple + GPT-4o Judge	4.2560	4.2445

Table 2: Official LLM-as-a-Judge performance on the FinCausal 2026 Blind Test Set (Scored 1 to 5). The ablations indicate that inference-only repetition and multi-stage pipelines (Ensembles, RAG, Correctors) generally resulted in lower scores compared to single-stage, natively fine-tuned models. (*) indicates the official submission.

potential conflict between structural training constraints and Reinforcement Learning from Human Feedback (RLHF). Heavily aligned models optimized for conversational naturalness may penalize or misinterpret highly unnatural, duplicated input structures during generative decoding. This indicates that structural prompting techniques effective on base-aligned open models may conflict with the safety alignment layers of proprietary models.

3. Training-Time vs. Inference-Time Repetition: To examine whether the performance gains stem from the training paradigm or merely from inference-time context duplication, we conducted a targeted ablation on the blind test set. When the proprietary model was fine-tuned on the *Simple* configuration but evaluated using the *Repeated* prompt during inference, its score decreased (from 4.7396 down to 4.6143 in Spanish). This supports the idea that forcing prompt repetition solely at inference introduces out-of-distribution formatting noise, and that the model benefits from internalizing the attention mechanism during the SFT phase.

4. The Limitations of Multi-Stage Pipelines: Recent NLP extraction tasks frequently deploy multi-stage pipelines. However, our ablations (Table 2) suggest these architectures can be less effective for strict verbatim extraction. Injecting dynamic context via Retrieval-Augmented Generation (RAG, utilizing 3-shot semantic retrieval for in-context examples) introduced external noise, slightly lowering the score. Similarly, utilizing a powerful meta-model (`GPT-4o`) as a post-generation corrector via strict formatting prompts caused a decrease in performance (from 4.7040 to 4.2560). Qualitative analysis indicated that the corrector model prioritized grammatical completeness over verbatim copying, disrupting the required span boundaries. Furthermore, Ensemble Voting (majority consensus across

5 high-temperature generations) introduced token-level variations that diluted the Exact Match consistency.

6. Error Analysis and Qualitative Comparison

To understand the mechanism by which Prompt Repetition improves Exact Match (EM) scores, we conducted a qualitative analysis of the residual errors in the baseline models. As demonstrated in Table 4 (Appendix C), the generative formulation introduces specific hallucination patterns.

For instance, consider a boundary error hallucination where the baseline *Simple Prompt* misses the exact starting boundary by appending an introductory connector (e.g., extracting "***As** external threats become more sophisticated*" instead of "*external threats become more sophisticated*"). Furthermore, when faced with causal chains, the baseline model occasionally truncates the extraction, or exhibits generative stutters (Example 1) and mid-generation stops (Example 4).

As shown in Table 4 (Appendix C), internalizing the *Repeated Prompt* during Supervised Fine-Tuning acts as an attention constraint, encouraging the model to respect verbatim spans and reducing these generative tendencies across the tested open-weights architectures. The high verbatim copy-rate achieved by the Repeated configuration supports our hypothesis that the SFT process enforces a unidirectional "cross-reading" mechanism.

7. Conclusion

This study investigated the effectiveness of a structural training constraint combining Supervised Fine-Tuning with prompt repetition for strict span extraction. We demonstrated that extending prompt repetition from an inference-time heuristic to a training-time supervision approach acts as a context anchoring mechanism, allowing generative causal LLMs to internalize cross-reading behaviors. The results suggest that structurally simple, natively fine-tuned generative models can perform reliable verbatim extraction, neutralizing language-specific biases to achieve balanced multilingual performance without relying on complex multi-stage pipelines. Although evaluated in the financial domain, the proposed structural constraint may extend to other high-precision extraction tasks where strict verbatim copying is required.

8. Limitations and Future Work

Our approach demonstrates practical empirical performance but also highlights several avenues for future research. The models were fine-tuned on a relatively small, domain-specific dataset (3,600 bilingual samples), which was effective in this context; however, extending structural training constraints to larger, multi-domain corpora would help assess broader generalization. While our study focused on the financial sector, adapting prompt repetition for verbatim extraction in other domains, such as biomedical relation extraction or legal evidence analysis, remains an open challenge.

The LLM-as-a-judge metric provides a useful assessment of answer adequacy but may introduce implicit alignment biases. Future work could explore hybrid evaluation frameworks to better relate these subjective judgments to deterministic metrics like Exact Match.

Computationally, prompt repetition imposes only minor overhead. As shown in Table 3 (Appendix B), it increased training time by 0.02 hours, added a negligible $+0.0009$ kg of CO_2 emissions, and caused only a slight rise in peak VRAM (38.67 GB vs. 39.31 GB), indicating feasibility for mid-scale models.

Finally, while our results show a statistically significant improvement in Exact Match scores (+3.5%), achieving absolute boundary determinism remains a challenge for generative architectures, even under strict SFT constraints. Our interpretation that prompt repetition may promote a unidirectional “cross-reading” mechanism is supported by behavioral evidence, such as reduced generative stutters and boundary offsets. Nonetheless, a formal extraction and visualization of multi-head attention maps could provide further validation and represents a promising direction for future interpretability research.

9. Acknowledgements

We thank the organizers of the FinNLP and FNP workshops for their dedication to the financial NLP community. We acknowledge the creators of the FinCausal 2026 dataset (Moreno-Sandoval et al., 2026), whose bilingual corpus enabled the cross-lingual analyses presented in this research.

10. Bibliographical References

Tom B Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

Sohom Ghosh and Sudip Kumar Naskar. 2022. Lipi at fincausal 2022: Mining causes and effects from financial texts. In *Proceedings of the 4th Financial Narrative Processing Workshop*, pages 121–123.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves large language models. *arXiv preprint arXiv:2512.14982*.

Zhiheng Lyu et al. 2022. Dcu-lorcan at fincausal 2022. In *Proceedings of the 4th Financial Narrative Processing Workshop*.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The financial document causality detection shared task (fincausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination in causal q&a: The efficacy of fine-tuning over prompting in llms. In *Proceedings of the Joint Workshop of the 9th FinNLP, 6th FNP, and 1st LLMFinLegal*, pages 253–258.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157.

Anik Saha, Jian Ni, Oktie Hassanzadeh, et al. 2022. Spock at fincausal 2022: Causal information extraction using span-based and sequence tagging models. In *Proceedings of the 4th Financial Narrative Processing Workshop*, pages 108–111.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and S.R. Balasundaram. 2025. Sarang at fincausal 2025: Contextual qa for financial causality detection combining extractive and generative models. In *Proceedings of the Joint Workshop of the 9th FinNLP, 6th FNP, and 1st LLMFinLegal*, pages 242–247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

11. Language Resource References

Language Resources

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

A. Prompt Templates

We detail the exact prompt structures utilized during both Supervised Fine-Tuning and inference.

Simple Prompt (Zero-Shot SFT Baseline):

```
System:
You are a financial expert participating in FinCausal 2026.
Task: Extract the exact cause or effect from the provided financial text.
Rules:
1. The answer must be a VERBATIM extraction.
2. If the text contains a complex causal chain,extract the full relevant se-
quence.
3. Do not add introductory words.

User:
Context: {context}
Question: {question}
```

Repeated Prompt (Cross-Reading Configuration):

```
System:
You are a financial expert participating in FinCausal 2026. Task: Extract
the exact cause or effect from the provided financial text.
Rules:
1. The answer must be a VERBATIM extraction.
2. If the text contains a complex causal chain,extract the full relevant se-
quence.
3. Do not add introductory words.

User:
Context: {context}
Question: {question}
Context: {context}
Question: {question}
```

B. CodeCarbon Environmental Tracking

Model	Setting	Time (h)	CO_2 (kg)	Peak VRAM (GB)
Qwen2.5-14B-Instruct-1M	Simple	0.20	0.0051	38.67
Qwen2.5-14B-Instruct-1M	Repeated	0.22	0.0060	39.31

Table 3: Empirical computational cost and environmental impact for fine-tuning on a single NVIDIA L40S (48GB) GPU. Total training time, peak VRAM utilization, and CO_2 emissions were directly tracked using the CodeCarbon library. Measurements were taken during a dedicated reproducibility run using identical hyperparameters, data, and hardware.

C. Qualitative Examples

Target Span	Simple Prompt Baseline	Repeated Prompt
Type 1: Boundary Offset & Over-generation		
[Qwen2.5-14B-Instruct-1M - EN] the £9.4m improvement in underlying operating cash flows	the £9.4m improvement in underlying operating cash flows **offset by a £2.0m increase in outflows...**	**the £9.4m improvement in underlying operating cash flows**
[GPT-4.1-Nano - EN] external threats become more sophisticated, and the potential impact of service disruption increases	**As** external threats become more sophisticated, and the potential impact of service disruption increases	**external threats become more sophisticated, and the potential impact of service disruption increases**
Type 2: Generative Stutter & Typographical Drops		
[Qwen2.5-7B-Instruct-1M EN] Our business serving the grocery sector benefited from several new accounts although the additional business won, combined with a competitive marketplace	Our business serving the grocery sector benefited from several **new new** accounts	**Our business serving the grocery sector benefited from several new accounts although the additional business won, combined with a competitive marketplace**
[Qwen2.5-14B-Instruct-1M - ES] no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de deuda	no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de **deud**	**no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de deuda**
Type 3: Truncation (Premature Stops)		
[GPT-4.1-Nano - ES] parámetros como el suministro de materias primas, la utilización de los quemadores, los sensores instalados o el balance entre energía fósil y eléctrica pueden ser gestionados de una manera más rápida, moderna y eficiente	**parámetros como el s**	**parámetros como el suministro de materias primas, la utilización de los quemadores, los sensores instalados o el balance entre energía fósil y eléctrica pueden ser gestionados de una manera más rápida, moderna y eficiente**
[Qwen2.5-14B-Instruct-1M - EN] The key partnerships established with leading European manufacturers	**The key partnerships**	**The key partnerships established with leading European manufacturers**

Table 4: Qualitative comparison of extraction errors. A focused evaluation of six samples illustrates how Prompt Repetition reduces generative stutters, boundary misalignments, and premature truncations across multiple architectures without requiring an external corrector model.

LeedsMEng26: Qwen + Gemini for FinCausal 2026 Causality Detection in Financial Narrative Texts

Ayomide Ivienagbor*, **Idrees Asad***, **Rijul Shrestha***
Yasemin Bal*, **Zahaab Nadeem***, **Zaid Shahrouri***
*University of Leeds, Leeds, United Kingdom
sc22ai@leeds.ac.uk, sc22i2a@leeds.ac.uk, sc22r2s@leeds.ac.uk,
sc22y2b@leeds.ac.uk, sc22zn@leeds.ac.uk, sc21z2s@leeds.ac.uk

Abstract

This paper presents the LeedsMEng26 system for the FinCausal 2026 shared task Moreno-Sandoval et al. (2026a) on financial causality detection in narrative texts. The task is formulated as extractive question answering over English and Spanish financial reports, where systems must return a verbatim span from the context that answers an abstractive question about a cause or an effect. We propose a two-stage pipeline consisting of candidate span generation followed by span verification and boundary refinement under a strict extractiveness constraint. We evaluate both an extractive RoBERTa-based baseline and instruction-tuned large language models. Results show that Qwen-2.5-1.5B-Instruct is a stronger candidate generator than the RoBERTa baseline, and that a second-stage verifier further improves answer boundary accuracy and overall adequacy. Our best configuration, Qwen-2.5-1.5B-Instruct with Gemini-2.5-flash refinement, achieved an adequacy score of 4.7000 for English and 4.6143 for Spanish. These findings suggest that a modular generation-and-verification pipeline is effective for extractive financial causality detection.

Keywords: Financial Narrative Processing, Causality Detection, Natural Language Processing, Question Answering, Multilingual NLP, Financial Text Analytics

1. Introduction

Financial narratives such as annual reports, earnings announcements, management commentaries, and regulatory filings play a central role in how firms communicate with investors, regulators, and the public. These documents often describe not only what has happened but also why it occurred. They do this through explicit or implicit causal statements, such as linking changes in performance to macroeconomic events or strategic decisions. Understanding these causal links is essential for interpreting corporate performance, assessing risk, and forming expectations about future developments. Manually analysing narratives at scale is costly and time-consuming, particularly given the volume and growing complexity of financial disclosures. This has motivated increasing interest in applying natural language processing (NLP) to financial text.

Within this emerging area, financial causality detection focuses on identifying cause–effect relations in financial documents. Automatically extracting such relations can support tasks such as risk analysis, fraud detection, forecasting, and explainable decision support. It also enhances transparency by making implicit reasoning patterns in corporate communication more explicit and machine-interpretable. Despite these benefits, financial text presents several challenges: it is often technical, domain-specific, and highly contextual, and causal language may be expressed in subtle, indirect or multi-sentence forms. Moreover, financial narratives frequently combine quantitative data with qualitative explanations, requiring models to integrate numerical reasoning with linguistic understanding. Recent shared tasks, including the FinCausal track at the Financial Narrative Processing (FNP) workshop, provide benchmark datasets and evaluation protocols for studying these problems in a controlled setting, thereby enabling systematic comparison of modelling approaches.

This paper describes the LeedsMEng26 system for the FinCausal 2026 shared task (Moreno-Sandoval et al., 2026b), which formulates financial causality detection as question answering over English and Spanish financial texts. Given an abstractive question and a short context paragraph, the system must return an extractive span from the context that answers either the cause or the effect. We study both extractive and instruction-tuned generative approaches under a strict extractiveness constraint, and propose a two-stage pipeline: (i) candidate span generation and (ii) span verification and boundary refinement using a verifier LLM constrained to copy a contiguous substring.

2. Related Work

The FinCausal shared tasks have progressively advanced research on causality detection in financial narratives, moving from span extraction in earlier editions toward question-answering and more generative evaluation settings in recent years. The current edition, FinCausal 2026, continues the multilingual English–Spanish setting and retains the use of annotated contexts paired with abstractive questions and extractive answers. However, it introduces two notable changes: a new random parti-

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 160–168

tioning of the 2026 dataset and an LLM-as-a-judge evaluation metric, which scores responses on a 1–5 adequacy scale and replaces the previous Semantic Answer Similarity (SAS) and Exact Match (EM) metrics used in 2025 (Moreno-Sandoval et al., 2025).

Earlier editions focused more on span- and token-level causality extraction. FinCausal 2023 expanded the task across English and Spanish, using span-level Exact Match (EM) and token-level weighted F1 for evaluation, while encouraging multilingual and prompt-based approaches (Moreno-Sandoval et al., 2023). FinCausal 2022 focused on causality in quantified financial facts, with the winning SPOCK system using an ensemble of RoBERTa-Large sequence-tagging models with the BIO scheme (Mariko et al., 2022). Overall, the FinCausal series shows a progression from structured cause–effect extraction to multilingual reasoning, QA-based formulations, and more flexible generative evaluation.

(Moreno-Sandoval et al., 2025) introduces FinCausal 2025 competition entries and it was used as a guide to see where the most recent advancements for the task were. Participants adopted a range of approaches, spanning discriminative extractive QA models, generative LLMs with prompt engineering (simple, CoT, few-shot), and varying use of fine-tuning and quantization. Notably, strong scores were achieved even without fine-tuning in some cases, highlighting that fine-tuning was not the only route to competitive performance.

The (Trivedi et al., 2025) system paper was useful for our work because it provided a strong, task-aligned example of a hybrid extractive to generative refinement pipeline (RoBERTa-based span prediction followed by Gemma2-9B refinement) that directly targets common QA boundary and coherence errors while remaining competitive on the official SAS/EM metrics.

3. Dataset and Task

The FinCausal 2026 task is formulated as a generative question answering problem over financial annual reports, where systems must identify either the cause or the effect corresponding to an abstractive question. We decided to participate in both sub-tasks, training models on both the English and Spanish texts.

The training data are drawn from the FinCausal 2026 dataset Moreno-Sandoval et al. (2026b), which is provided in CSV format, with each file containing 2000 rows. Each instance was in the following form: (i) **ID** - its unique identifier; (ii) a **context**, corresponding to a paragraph extracted from a financial report; (ii) an **abstractive question**, asking for either the cause or the effect of a described event; and (iii) an **answer**, which is an extractive span taken precisely from the context. Although the question is abstractive, the expected output is a text span grounded strictly in the provided paragraph.

The 2026 edition introduces a revised and expanded dataset, including more complex causal structures and rephrased questions designed to require deeper reasoning. The training and test sets are randomly partitioned from this updated dataset. The task focuses on explanatory cause–effect relations within the text, especially where specific events result in measurable financial outcomes.

For our submission, we approached the problem from both an extractive and a generative perspective. We first experimented with span-based extraction models to exploit the extractive nature of the gold answers, and subsequently explored fine-tuning and combining large language models to assess whether generative approaches could better capture complex causal structures.

4. Methodology

4.1. Task Formulation

Financial causality extraction is approached as an extractive question answering (QA) task. Each instance consists of a financial context c and a question q specifying a causal relation. The system produces an answer span a that must appear verbatim within c. We apply this extractive constraint strictly to all system variants. This ensures that predictions are always drawn from the provided context.

We report span-level Exact Match (EM) and Semantic Answer Similarity (SAS) metrics. EM evaluates strict boundary correctness. SAS accounts for minor boundary deviations. The official leaderboard score is also tracked where applicable.

4.2. Pipeline Overview

The proposed approach utilises a two-stage framework to enhance boundary accuracy and preserve the extractive constraint:

1. **Candidate generation:** an extractive QA model produces a candidate span $\hat{a}$ from the context.

2. **Span verification and refinement:** an instruction-following LLM receives $(c, q, \hat{a})$ and either confirms $\hat{a}$ or corrects span boundaries, while being constrained to return a verbatim substring of c.

This design distinctly separates relevant evidence retrieval in stage 1 from boundary correction

and consistency verification in stage 2. Preliminary error analysis revealed that boundary truncations, such as missing causal qualifiers, and boundary overruns, such as the inclusion of adjacent clauses, are the most frequent failure modes in pure extractive models. The refinement stage specifically addresses these errors while maintaining strict adherence to the input context.

Figure 1 illustrates the two-stage candidate generation and refinement pipeline used in our system.

All systems employ a unified preprocessing & post-processing pipeline implemented using the Hugging Face `transformers` library. Inputs are constructed by concatenating the question and context with the model-specific tokeniser. For extractive QA models, start and end position labels are derived from the gold answer span.

Predicted spans undergo lightweight normalisation as follows: (i) whitespace trimming and deduplication, and (ii) minor punctuation trimming at span boundaries where it does not alter content. The extractive constraint is enforced through substring verification, requiring the produced answer to appear as a contiguous substring in the given context. If a refinement model outputs text that violates this constraint, the system reverts to the pre-refinement candidate.

4.3. Baseline 1: Extractive QA with RoBERTa

The first baseline employs `question-answering-roberta-base-s-v2`, a RoBERTa-base encoder fine-tuned for extractive QA. Given (c, q), the model produces probability distributions over context tokens for the start and end indices of the answer span. The highest-scoring (i, j) pair is selected, and the corresponding substring is returned verbatim.

We fine-tuned the model on the official FinCausal English training data. We optimised based on the extractive QA objective, minimising cross-entropy loss over the gold start and end token positions. This baseline serves as a strictly extractive reference point with no generative components.

4.4. Baseline 2: RoBERTa + GPT Refinement

As a result, it was concluded that RoBERTa was sensitive to boundary errors that affected the consistency of meanings. Early truncation and overrun errors were the main issues; consequently, a second-stage refinement step is introduced using GPT-3.5 as a verifier.

The refiner is provided with the context, question, and candidate span predicted by RoBERTa. It is instructed to:

- Output *only* the final answer span (no explanation),
- Copy the span verbatim from the context (no paraphrasing),
- Keep the candidate unchanged if correct, and
- Correct boundary errors by expanding or contracting to the shortest correct substring when multiple valid spans exist.

The method ensures compliance with the context and achieves improved span boundary precision. If the refined span is not a substring of the context, the original candidate is used.

4.5. Baseline 3: Qwen-2.5-1.5B-Instruct as Candidate Generator

The third baseline replaces the RoBERTa extractor with Qwen-2.5-1.5B-Instruct, a decoder-only instruction-tuned LLM with multilingual capability. Unlike encoder-style QA models that predict token positions, Qwen generates text directly; therefore, constrained prompting is used to enforce extractive behaviour.

The candidate-generation prompt explicitly requires the model to output a verbatim substring from the context and nothing else. Despite these directives, we still observed boundary drift and occasional paraphrasing. These were similar to those seen with RoBERTa when the model attempted to be 'helpful.'

Low-Rank Adaptation (LoRA) is employed, updating only low-rank matrices inserted into selected attention and feed-forward layers while keeping the base weights frozen, enabling Qwen to adapt to the task with limited computational resources and resulting in reduced training memory and cost compared to full fine-tuning.

Reinforcement learning (RL) post-training is explored using a composite reward. The reward encourages extractive correctness and semantic fidelity. It combines: (i) Exact Match (EM), (ii) a semantic similarity component between the prediction and the gold span, and (iii) an auxiliary judge score reflecting span correctness and boundary precision. At inference time, we use conservative decoding (low temperature) to reduce variability and discourage paraphrasing. Qwen-2.5-1.5B-Instruct was also fine-tuned using Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. (Schulman and Lab, 2025)

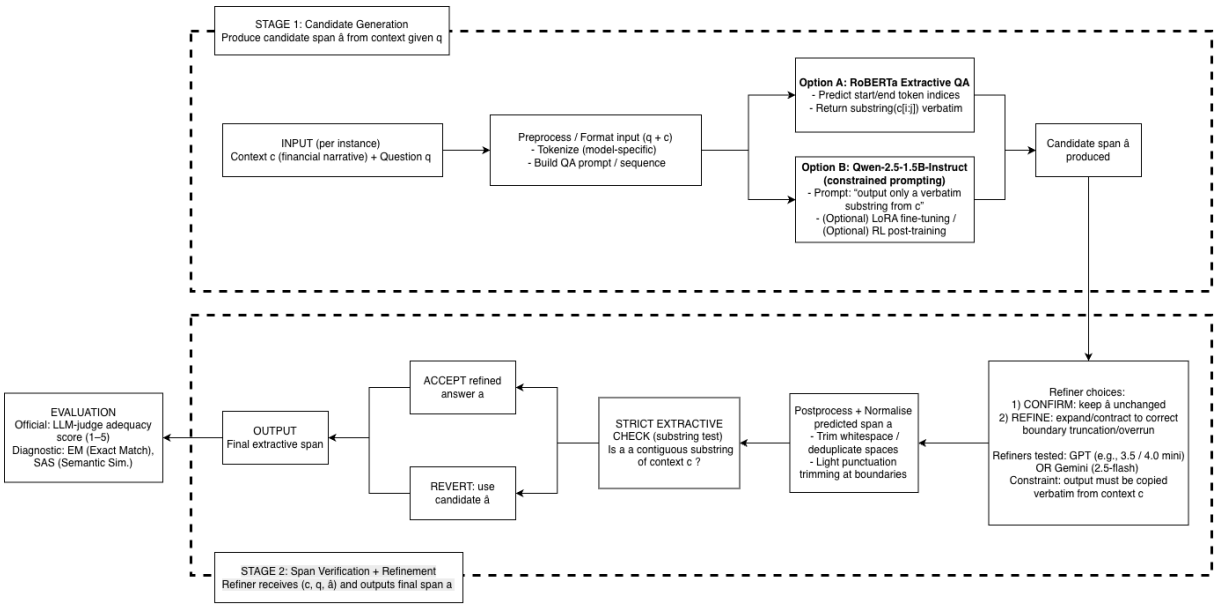


Figure 1: Two-stage pipeline for financial causality extraction.

4.6. Hybrid: Qwen + GPT Refinement

Qwen generates an initial candidate span under strict prompting, and the verifier model performs second-stage verification and boundary correction, replicating the two-stage process used in the RoBERTa + verifier configuration.

We perform substring verification after refinement as before. If GPT-based refinement (GPT-3.5 and GPT-4.0 mini) violates extractiveness, the system reverts to the original Qwen candidate. This hybrid approach uses Qwen's strong instruction-following together with GPT-3.5 and GPT-4.0 mini for verification, reducing hallucination and improving boundary precision. The following configuration further explores this two-stage strategy by integrating a different refinement model, as outlined next.

4.7. Final System: Qwen + Gemini Refinement

Ultimately, we used Gemini-2.5-Flash as the refinement model. Qwen generates a candidate span using strict extractive prompting, then provides the prediction to Gemini and is instructed to verify correctness and adjust the boundaries if needed.

Similar to previous hybrids, we apply minimal normalisation and substring verification, reverting to the candidate span if refinement breaks extractiveness. This configuration provided the most consistent boundary corrections and the strongest overall performance across English and Spanish settings. The next sections outline the experimental setup supporting these results.

4.8. Refinement Prompt

The refinement stage relies on a constrained prompt designed to correct boundary errors while preserving the extractive constraint.

The verifier model receives the context, question, and candidate span produced by Qwen and is instructed to return only the final corrected answer.

The prompt enforces several rules: (i) the output must be a verbatim span from the context, (ii) no explanations or additional text may be produced, (iii) the candidate span should be returned unchanged if it is already correct, and (iv) if multiple spans are possible, the shortest correct substring should be selected.

A shortened version of the prompt is shown below.

> You are correcting a short answer. Output only the final answer. Return a verbatim span from the context. Do not add explanations. If the answer is already correct, return it unchanged. If multiple spans are possible, choose the shortest correct span.

The full prompt, including the in-context examples used during inference, is provided in Appendix A. A Spanish version of the prompt with identical constraints was used for Spanish inputs.

4.9. Experimental Setup

- **Data:** Experiments used the FinCausal 2025 dataset. Training was performed using only the English data. Spanish examples were evaluated without additional training.

- **Data Split:** The English dataset consisted of 2000 instances, which were divided into 75% training and 25% validation data.
- **RoBERTa Training:** The parameters used were learning rate of 3×10^{-5}, weight decay 0.01, and a linear scheduler with warmup ratio 0.1. The batch size was 4 with gradient accumulation of 2 (effective batch size 8). The best model was selected based on validation loss.
- **Qwen Fine-tuning:** Qwen-2.5-1.5B-Instruct was fine-tuned using LoRA with rank $r = 16$, $\alpha = 32$, and dropout 0.0. Bias parameters were not adapted, and gradient checkpointing was enabled using Unsloth.
- **Decoding:** Low-temperature decoding was used during inference. Outputs were constrained to be substrings of the input context to ensure extractive answers.
- **Reproducibility:** Random seeds and preprocessing were kept consistent across all experiments.

5. Experiments and Results

5.1. Evaluation Protocol

The FinCausal 2026 shared task uses an LLM-as-a-judge evaluation protocol. Each system prediction is scored on a 1–5 adequacy scale. The judge's score rewards answers that fully address the question using evidence from the context. It penalises truncations, overruns, and non-extractive outputs.

In addition to the official adequacy score, we report **Exact Match (EM)** and **Semantic Answer Similarity (SAS)** as *diagnostic* metrics to analyse span boundary quality and semantic correspondence during development. Unless stated otherwise, EM/SAS are computed by comparing predictions against available labelled data and are used for internal evaluation rather than leaderboard ranking.

5.2. English Results

System Configuration	Score
RoBERTa + GPT-3.5	3.8200
Qwen-2.5-1.5B-Instruct + GPT-3.5	4.5080
Qwen-2.5-1.5B-Instruct	4.6060
Qwen-2.5-1.5B-Instruct + GPT-4.0 mini	4.6240
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	4.7000

Table 1: Best performing English submission per system configuration (LLM-judge adequacy score).

Table 1 summarises the best-performing English submission for each system configuration. Overall, performance improves monotonically as the candidate generator becomes more instruction-aligned and as verification-based refinement is introduced.

The RoBERTa-based hybrid baseline (**RoBERTa + GPT-3.5**) achieves an adequacy score of **3.8200**. While the extractive QA model reliably returns substrings from the context, qualitative inspection indicates that it repeatedly faces minor span boundary errors (e.g., missing causal qualifiers or including adjacent clauses), which reduces adequacy under judge-based scoring.

Replacing the encoder-based extractor with an instruction-tuned decoder model yielded substantial improvement. **Qwen-2.5-1.5B-Instruct** achieves **4.6060**, indicating that constrained, instruction-following generation is better aligned with the causality-oriented QA prompts and the judge's emphasis on completeness and adequacy.

Adding a second-stage verifier provides further gains. **Qwen + GPT-4.0 mini** reaches **4.6240**, indicating that refinement helps correct residual boundary variances and improves answer adequacy.

The best-performing English configuration is **Qwen + Gemini-2.5-flash**, achieving **4.7000**. This result supports the effectiveness of a generation verification pipeline in which a strong refiner corrects subtle span boundary errors while preserving the extractive constraint.

For diagnostic analysis, Table 3 reports EM and SAS on labelled English data. Both metrics increase consistently across configurations (EM from **0.3500** to **0.7415**; SAS from **0.8850** to **0.9460**), indicating that adequacy gains are accompanied by improvements in boundary accuracy and semantic correspondence, rather than reflecting superficial formatting differences.

5.3. Spanish Results

System Configuration	Score
RoBERTa + GPT-3.5	3.9264
Qwen-2.5-1.5B-Instruct + GPT-4.0	4.4692
Qwen-2.5-1.5B-Instruct	4.5030
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	4.6143

Table 2: Best performing Spanish submission per system configuration (LLM-judge adequacy score).

Spanish results (Table 2) follow similar trends, with the strongest performance again obtained by verification-based refinement using Gemini-2.5-Flash.

The RoBERTa + GPT baseline achieves **3.9264**. The standalone **Qwen-2.5-1.5B-Instruct** model im-

proves adequacy to **4.5030**, demonstrating strong cross-lingual generalisation in Spanish under extractive prompting.

Unlike in English, adding GPT-based refinement slightly reduced performance (**4.4692** vs **4.5030**), suggesting that the refiner may occasionally overcorrect boundaries or produce outputs that are less well aligned with the judge's adequacy preferences for Spanish instances. In contrast, **Qwen + Gemini-2.5-flash** yields the best Spanish score of **4.6143**, indicating that Gemini provides more reliable verification and boundary correction in the bilingual setting.

5.4. Overall Analysis

Across both languages, two observations are deduced. First, instruction-tuned candidate generation (Qwen) better aligns with the task format and produces more adequate spans with limited prompting. Secondly, a second-stage verifier further improves robustness using correcting boundary truncations and overruns, with **Gemini-2.5-flash** providing the most reliable refinement among tested models. RoBERTa achieved lower scores on both the English and Spanish tasks, indicating lower effectiveness than the alternative approaches evaluated.

Overall, the hybrid **Qwen + Gemini-2.5-Flash** system achieved the best results in both English and Spanish, indicating that generation verification pipelines are effective for extractive financial causality tasks evaluated using an adequacy-oriented LLM judge.

Model	EM Accuracy	SAS
RoBERTa + GPT-3.5	0.3500	0.8850
Qwen-2.5-1.5B-Instruct + GPT-3.5	0.6295	0.9155
Qwen-2.5-1.5B-Instruct	0.6855	0.9366
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	0.7415	0.9460

Table 3: Diagnostic EM and SAS results for English (computed on labelled data for development).

6. Strengths and Limitations

A key strength of our work is that, despite limited time and resources, we achieved strong competitive performance, placing 7th in the English category with only marginal differences from teams ranked above us. This result demonstrates that our approach was effective and robust even under practical constraints.

Our pipeline also benefits from a clear and structured design that combines the reliability of extractive question answering with the contextual reasoning benefits of instruction-tuned Large Language Models and verification/refinement steps applied after the initial prediction. By progressing from a strong extractive baseline (RoBERTa) to an instruction-following model (Qwen) and then adding refinement stages (GPT-3.5, GPT-4.0 mini, and Gemini-2.5-Flash), we were able to isolate which components contributed most to performance gains and reduce common QA errors such as span boundary drift. In particular, our two-stage candidate and verifier setup made the system more reliable. The verifier checks the initial span and fixes common issues such as answers being too short (truncated) or too long, while also ensuring the final output stays strictly extractive. This iterative setup also made our experiments more interpretable, since each stage had a clear and measurable role in improving Exact Match and semantic alignment.

Despite the effectiveness of our pipeline, several constraints limited the scope of our experiments and likely capped performance:

- **Time constraints due to schedule clashes.** The shared task timeline overlapped with our university exam period, causing us to begin experimentation later than planned and reducing the time available for broader hyperparameter searches and ablation studies.
- **Limited compute access.** We did not have access to the university's powerful GPUs, which restricted our ability to train larger models, run longer fine-tuning schedules, or explore more compute-intensive approaches. A single NVIDIA RTX 3050 was used for the reinforcement learning finetuning of Qwen-2.5-1.5B.
- **Restricted access to the newest proprietary LLM APIs.** We were unable to use the latest frontier LLM APIs. Access to stronger models for refinement and verification could plausibly have improved boundary correction and reduced rare failure cases.
- **Training focused only on English.** Our main training and optimisation effort was concentrated on the English dataset rather than fully training separate systems for both English and Spanish. This likely reduced Spanish performance relative to what could be achieved with language-specific fine-tuning and validation.

7. Dataset Feedback

One limitation of the dataset is its relatively small size, which limited the amount of supervision avail-

able for adapting the models to the task. While extractive models remained reasonably stable, larger generative models were more sensitive to the limited supervision and showed less consistent performance. This limitation influenced our decision to adopt a multi-stage approach with a separate verification step, rather than relying solely on direct fine-tuning. A larger training set would likely allow more effective fine-tuning of generative models and lead to more stable and reliable results overall.

8. Conclusion and Future Work

Across the project, we demonstrate that a modular extractive QA pipeline is an effective approach for financial causality extraction, achieving a 7th-place ranking in the English track with only marginal differences from higher-ranked teams. Starting from an extractive baseline (RoBERTa) and progressively incorporating instruction-following modelling (Qwen) and refinement stages (GPT-3.5, GPT-4.0 mini, and Gemini-2.5-Flash), we achieved consistent improvements while keeping outputs strictly extractive. In particular, the candidate-verifier design helped reduce span boundary drift by correcting answers that were too short (truncated) or too long, and the staged structure made it clear which components were responsible for the improvements in Exact Match and semantic alignment.

For future work, this system will serve as a foundation for a Masters-level group project and will be expanded in both scope and capability. We aim to evaluate whether the approach generalises beyond financial reports to other domains such as medical or construction text, and to train and fine-tune stronger models using improved computational resources to further boost performance.

References

Dominique Mariko, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022*, pages 105–107, Marseille, France. European Language Resources Association.

Antonio Moreno-Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026a. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (fincausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, Sorrento, Italy. IEEE.

Moreno-Sandoval, Antonio and Torterolo Orta, Yanco Amor and Stanescu, Maria Alexia and Chatzi, Melina. 2026b. *The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset*. e-cienciaDatos.

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and S. R. Balasundaram. 2025. Sarang at FinCausal 2025: Contextual QA for financial causality detection combining extractive and generative models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 242–247, Abu Dhabi, UAE. Association for Computational Linguistics.

A. Appendix

A.1. English Refinement Prompt

The following prompt was used in the second-stage refinement step of the Qwen + Gemini pipeline.

You are correcting a short answer. **Rules:**

- Output **only** the final answer.
- Do **not** add explanations.
- Return a **verbatim span** from the context.
- If the answer is already correct, return it unchanged.
- Prefer the full sentence containing the answer if needed.
- If multiple spans are possible, choose the shortest correct one.
- Goal: maximise Exact Match.

Examples:

Context: Additionally, during the year the Group commenced work, under its EPCC contract with CGE, on four biogas-based power generation plants. As previously stated, due to financial constraints, progress was slower than initially expected and work has been temporarily suspended, awaiting the Company finalising an arrangement with CGE.

Question: What did financial constraints bring about?

Answer: progress was slower than initially expected and work has been temporarily suspended, awaiting the Company finalising an arrangement with CGE

Context: The Directors believe this growth is driven by consumer preferences moving away from chain and branded pubs and towards pubs with an individual identity and an ambience which reflects the local market.

Question: What explains the growth, according to the Directors?

Answer: consumer preferences moving away from chain and branded pubs and towards pubs with an individual identity and an ambience which reflects the local market

Context: The nature of the Group's operations creates an ongoing demand for fuel and therefore the Group is exposed to movements in market fuel prices. The Group enters into commodity derivative instruments to hedge such exposure where it makes commercial and economic sense to do so.

Question: What explains why the Group is exposed to movements in market fuel prices?

Answer: The nature of the Group's operations creates an ongoing demand for fuel

A.2. Spanish Refinement Prompt

For Spanish inputs, a language-adapted version of the refinement prompt was used to ensure consistent instruction following. The structure and constraints remained identical to the English prompt, but the instructions and examples were provided in Spanish.

Estás corrigiendo una respuesta corta.

Reglas:

- Devuelve **solo** la respuesta final.
- No añadas explicaciones.
- La respuesta debe ser un fragmento **verbatim** del contexto.
- Si ya es correcta, devuélvela igual.
- Si es necesario, prefiere la oración completa que contiene la respuesta.
- Si hay múltiples opciones, elige la más corta correcta.
- Objetivo: maximizar Exact Match.

Ejemplos:

Contexto: Debido a restricciones financieras, el progreso fue más lento de lo esperado y el trabajo ha sido suspendido temporalmente hasta que la empresa finalice un acuerdo.

Pregunta: ¿Qué provocaron las restricciones financieras?

Respuesta: el progreso fue más lento de lo esperado y el trabajo ha sido suspendido temporalmente hasta que la empresa finalice un acuerdo

A.3. Model Versions

Table 4: Model versions used in the experiments.

Model	Version / Checkpoint	Date Used
Qwen	Qwen2.5-1.5B-Instruct	March 2026
RoBERTa QA	question-answering-roberta-base-s-v2	March 2026
GPT-3.5	GPT-3.5	March 2026
GPT-4	GPT-4.0	March 2026
Gemini	Gemini 2.5 Flash	March 2026

Financial Causal QA via Instruction and Prompt Tuning of Gemma3-12B

Avinash Trivedi, Chindukuri Mallikarjuna
SRM University-AP,
Amaravati, Andhra Pradesh 522240, India
avinashtrivedi.2008@gmail.com

Abstract

In this paper we present a novel methodology that harnesses the power of prompt tuning applied directly to Gemma3-12B, a state-of-the-art generative large language model to enhance performance on complex natural language processing challenges. Instead of relying solely on extensive retraining, our approach leverages carefully crafted input prompts to steer the pre-trained Gemma-12B towards generating outputs with superior contextual accuracy and interpretability. Our experimental evaluation employed a composite LLM Score metric that quantifies both semantic coherence and relevance; under this framework, our system (Team Name: *Sarang*) achieved a score of 4.54, ranking 9th in the shared task. Furthermore, in the competitive task evaluation, our method demonstrated the potential of prompt tuning as a viable alternative to traditional fine-tuning approaches. This study not only demonstrates the practical benefits of integrating prompt engineering with large language models but also opens avenues for future research aimed at further optimizing model performance in domain-specific applications.

Keywords: FinCausal, Prompt Tuning, LLM Score

1. Introduction

Natural question answering systems which are expected to facilitate decision-making and market research should have a sophisticated sense of cause-and-effect structure embedded in financial narrative. Annual reports, earnings statements, etc., often describe events, antecedents and the consequences of those events such that it requires a keen sense of cause and effect. These linkages are computationally intensive and impractical to extract manually when faced with the large volume of modern financial text corpora. Therefore, automated causal question-answering reduces such limitations, simplifying the processing of financial information and making the produced financial intelligence more readable and understandable. It is on this basis that the creation of systems that can respond to causal questions derived through financial narratives has become one of the critical research areas.

The FinCausal 2026 Shared Task (Moreno-Sandoval et al., 2026a), which is structured into the Financial Narrative Processing Workshop, is focused on the improvement of the methodologies of causal question-answering in financial texts. The dataset has been carefully designed in terms of triadic settings of context, inquiry, and response whereby the participants are required to make an inference of the missing causal determinant on a financial passage and the query related to that passage. These interrogatives are highly abstracted and hence compel the examinees to identify antecedent or consequent items whereas the intended responses are specific extractive spans obtained out of the contextual contents. Therefore, the construct can be described as a graceful combination of classical span extraction and question answering with reasoning, and, as a consequence, such an elevated standard of subtle understanding of financial stories. Although modern large language models have made significant progress, it is still an extremely challenging task to identify causal relationships in the financial discourse: causes and effects are incorporated in a systematic way in such texts, specialised terminological registers are used, and long-term causal relationships are cultivated. The 2026 update makes the complexity intrinsic, with increased granularity of causal processes, and re-organization of ways of inquiry, to require a high level of inferential skill. This means that researchers have to come up with mechanisms that can be used to traverse complex causal networks, beyond surface pattern recognition. What adds to these inherent difficulties is the addition of a new judgment measure based on the use of Large Language Models as adjudicators (LLM-as-a-judge). This change of paradigm shifts the focus of emphasis on specific span recall to a general evaluation of the semantic fidelity and reasoning profundity. Correspondingly, analysis will be based on the ability of a system to encode cause-and-effect relationships with probable coherence, and not merely match annotated spans.

2. Literature study

The methodic identification and parsing out of causal relationships in the financial discourse has

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 169–174

become a critical project in the context of financial natural language processing. The FinCausal shared tasks have provided a significant impact on this pattern, providing strict guidelines and increasing complex evaluation models. In 2020, the first FinCausal attempt was launched, introducing the first annotated corpus on financial causality detection, which defines two fundamental subtasks: sentence-level classification and definition of cause-effect spans(Mariko et al., 2020) In 2021 and 2022, further cycles of improvement were made, along with annotation guidelines, increasing the data coverage, and advancing more complex causal patterns, including quantified facts and transformation-based relations (Mariko et al., 2021, 2022). These initial implementations demonstrated the effectiveness of transformer-based encoders in combination with highly structured span-extraction systems, and at the same time, the difficulty in representing implicit causality and cross-sentence reasoning. Based on this background, the 2023 version of FinCausal expanded its criterion to include multilingual tracks and redefined the role of the span-oriented extraction as the part of question answering or sequence generating (Moreno-Sandoval et al., 2023). The 2025 version also better specified its aims by incorporating a multilingual causal question-answer model that may be assessed by Exact Match (EM) and Semantic Answer Similarity (SAS) scores (Moreno-Sandoval et al., 2025). This redefinition marked the end of pure extraction and an incorporative reasoning and generation of answers. FinCausal is closely related to the development of financial question answering research, which places greater emphasis on financial reasoning and multi-step inference, at the cost of more traditional financial reporting (Chen et al., 2021). Taken together, these changes reflect a larger change in causal recognition, that is, surface-level identification to a more abstract financial reasonability that incorporates text with quantitative information. The effectiveness of hybrid architectures which combine extractive precision and generative reasoning is further supported by the recent literature. (Pilault et al., 2020) proposed a model in which an extractive item picks the relevant evidence to modulate a transformer-based generative model and thus showing a better contextual consistency. According to (Luo et al., 2022), extractive and generative QA models have been systematically compared, with the first type proving much more successful in generalisation in limited settings, while the latter type possesses the benefits of abstraction. The NeurIPS EfficientQA competition (Min et al., 2021) unveiled the trade-off between the computation efficiency and results, where well optimised lightweight extractive models can be competitive with state-of-the-art results. Basic transformer models like RoBERTa (Liu et al., 2019) have also enhanced the abilities of domain adaptation.

3. Dataset for FinCausal2026

This paper uses English version of FinCausal 2026 Question Answering dataset (Moreno-Sandoval et al., 2026b) published as part of FinCausal shared task, which is dedicated to detecting causal links in financial narratives. The task is expected to test systems, which are able to read financial texts and identify the cause or effect of financial events. The dataset is assembled out of the financial reports and corporate disclosures in which causal relations are often found in the descriptions of the company performance, market trends and economic situation. The dataset consists of 2000 training instances and 500 instances for testing. In the training dataset each instance includes *ID*, *Context*, *Question* and *Answer*. Whereas test data contains same attributes except *Answer*.

Considering a financial context and a causal question, the goal is to derive the right cause or effect.

4. Methodology

4.1. Few-shot prompting and Finetuning of Language Model

As a baseline, We tried few-shot prompting of various LLMs. Later started the finetuning of *consciousAI/question-answering-roberta-base-s*, then changed the checkpoint to *deepset/deberta-v3-large-squad2* followed by prompt based enhancement steps as in Fig 1, inspired from (Trivedi et al., 2025) to improve the response received from finetuned model. This technique was giving LLM score of 4.424. We also tried prompt tuning and from there we found our best performing system discussed in section 4.2.

4.2. System Submission

The current section outlines the proposed structure of the FinCausal Question Answering (QA) task. Fig 2 representing the architecture of the model submission. The current research involves a methodology that integrates the few-shot learning, automated prompt optimization and a large language model to identify causal relationships in finance in a systematic manner. The pipeline includes four main elements, dataset utilisation, few-shot prompt construction, prompt tuning, and model inference, which together allow performing sound causal reasoning on financial texts.

Prompt

```
{"role": "system",
"content": "You are a helpful assistant
that provides accurate and improved an-
swers."},
{"role": "user",
"content": """You are given a Context, a
Question, and an Answer.
1. If the Answer is 100% correct and is ex-
tracted verbatim from the Context, return
the exact same Answer.
2. If the Answer is incorrect or not fully
extracted from the Context, return an im-
proved version of the Answer that is ex-
tracted verbatim from the Context.
Context: {context}
Question: {question}
Answer: {answer} """}
```

Figure 1: Prompt for enhancement step

4.2.1. Few-Shot Prompt Construction

Few-shot learning has proven to be effective in instructing large language models to perform specialised reasoning (Brown et al., 2020; Wei et al., 2022). In the current methodology, few instances of the dataset are integrated into prompts as demonstrations. Each instance in the dataset is characterized by a financial context, a causal question, and the answer that clearly outlines the cause effect relationship. Such demonstrations offer implicit information to the language model, and it is able to identify how causal relationships are formulated in financial texts and how the appropriate responses can be produced.

4.2.2. Prompt Optimization using MIPROv2

To further enhance timely efficacy, the system will use MIPROv2 prompt optimization through the DSPy teleprompter framework (Khattab et al., 2022, 2024). DSPy provides a programmatic interface that is structured in such a way that it can be optimized and interactions with language models co-ordinated. The MIPROv2 teleprompter automatically optimizes prompts by sequentially sampling a space of instruction and few-shot examples. Through iterative evaluation, the system will pick highly effective prompts, thus improving the ability of the model to identify causal relationship and provide accurate answers for FinCausal QA task.

4.2.3. Model Integration and Inference

Gemma3:12B large language model(Gemma and DeepMind, 2024) is the refined version that performs the prompts and provides the strong language understanding and reasoning skills. The implementation of the model makes use of DSPy, which is used together with Ollama to enable efficient local execution and enable a controlled interaction with the model. In the process of inference, the system obtains a context and a causal query. The optimized prompt along with the selected few-shot examples are sent to the language model using the DSPy framework. The model then performs contextual reasoning on the financial text and then gives out the final answer thus discovering the relevant causal relationship.

Our final model was build on prompt tuning of Gemma3-12B using MIPROv2 of DSPy (Khattab et al., 2022, 2024). The tuned system prompt is in Fig 3 and best parameters are listed in Table 1

Hyperparameter	Value
auto	medium
max_bootstrapped_demos	10
max_labeled_demos	10
model	gemma3:12b

Table 1: MIPROv2 Hyperparameters

5. Experimental Results

The results of our experiments few-shot prompting, prompt tuning and including finetuning of *consciousAI/question-answering-roberta-base-s* and *deepset/deberta-v3-large-squad2* are listed in Table 2.

Technique	Model	LLM Score
Finetuning	roberta based	4.136
Finetuning + Enhancement	roberta based	4.288
Finetuning	deberta based	4.292
Finetuning + Enhancement	deberta based	4.302
Few-shot	gemma2:latest	4.424
Few-shot	qwen3:latest	4.418
Few-shot + light optimization	gemma3:12b	4.448
Few-shot + medium optimization	**gemma3:12b**	**4.54**

Table 2: Performance comparison on test set

The experimental findings prove the existence of a performance improvement between the conventional transformer-based fine-tuning strategies and LLM based few-shot strategies augmented with prompt tuning. First, both the RoBERTa-based and DeBERTa-based models perform competitively among the fine-tuning methods. The initial fine-tuned RoBERTa model has a score of 4.136 that is enhanced to 4.288 using prompt based response enhancement step mentioned in Fig 1. In the same manner, the fine-tuned DeBERTa model achieves 4.292, which is slightly higher than RoBERTa and it reaches 4.302 after applying enhancement step. These outcomes demonstrate the idea that the architectural variation (DeBERTa versus RoBERTa) comes with minor benefits, whereas the enhancement strategies are both able to provide incremental improvements to both models.

Conversely, the few-shot LLM-based models have a visible lead over all fine-tuned encoder mod-

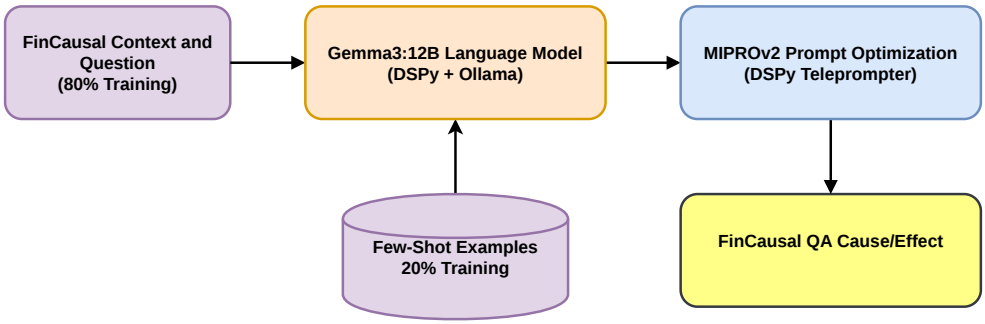


Figure 2: Model architecture

Prompt

```
Your input fields are:
1. context (str): Financial report excerpt
containing the causal relation
2. question (str): Question asking for the
cause or effect

Your output fields are:
1. answer (str): Exact causal span copied
from the context

All interactions will be structured in the
following way, with the appropriate values
filled in.

[[ ## context ## ]]
{context}

[[ ## question ## ]]
{question}

[[ ## answer ## ]]
{answer}

[[ ## completed ## ]]

In adhering to this structure, your objec-
tive is:

You are a financial causal reasoning expert.

The answer to the question (cause or effect)
is ALWAYS explicitly stated in the provided
context.

Extract the exact text span from the context
that answers the question.

Rules:
- The answer MUST be copied verbatim from the
context.
- Do NOT paraphrase.
- Do NOT add any extra words.
- Return only the precise causal phrase.
```

Figure 3: Tuned system prompt

els. The few-shot set up using Gemma2 also gives 4.424, and Qwen3 gives 4.418, indicating the high level of generalisation of large instruction-tuned models without task-specific fine-tuning. This indicates that prompt-based learning on modern LLMs performs better at this task environment than in traditional fine-tuning. Lastly, better performance is further improved by optimisation of bigger LLMs. The few-shot plus light optimisation experiment with Gemma3 (12B) has a result of 4.448 and medium optimisation has the most optimal result of 4.54, which is the best overall result. The development of this process underlines the fact that prompt tuning methods significantly enhance the efficiency of a model. Overall, the outcomes show that there is a definite trend: bigger LLMs with organised prompt tuning outperform classic fine-tuned transformer baselines, achieving the best results in an experimental environment.

6. Conclusions and Future Work

Within this investigation we have delineated several experimental paradigms, including few-shot learning, fine-tuning procedures, and prompt tuning. Our most effective submission i.e. Gemma3-12B prompt tuning achieved an LLM Score of 4.54.

Looking ahead, future work will concentrate on overcoming computational resource constraints, thereby permitting an in-depth exploration of prompt tuning strategies for larger LLMs. Furthermore, investigating data augmentation techniques to fine-tune the deberta based checkpoint represents another promising research direction. Finally, the potential integration of LLM agents remains a viable avenue for subsequent experimental inquiries.

7. Bibliographical References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over

financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3697–3711. Association for Computational Linguistics.

Team Gemma and Google DeepMind. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. 2022. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 7–22, Dublin, Ireland and Online. Association for Computational Linguistics.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Sewon Min et al. 2021. Neurips 2020 efficiency tqa competition: Systems, analyses and lessons learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *PMLR*, pages 86–111.

Antonio Moreno-Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (fincausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026a. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA.

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026b. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

Antonio Moreno-Sandoval, Jordi Porta Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (fincausal 2023). In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData 2023)*, pages 2855–2860, Sorrento, Italy. IEEE.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 9308–9319. Association for Computational Linguistics.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and SR Balasundaram. 2025.

Sarang at fincausal 2025: Contextual qa for financial causality detection combining extractive and generative models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 242–247.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

QRAFT: QLoRA Retrieval-Augmented Fine-Tuning for Causal Span Extraction in Financial Documents

Bavya Sarda[1], **Pulkit Chatwal**[2], **Sonal Dabra**[3]
[1]Galgotias University, Greater Noida, India
[2]Rajiv Gandhi Institute of Petroleum Technology, India
[3]Sony Research, India
{bhavyasarda19, pulkitchatwal, sonaldabral26}@gmail.com

Abstract

Understanding *why* financial outcomes occur is as important as knowing *what* they are. Annual reports and regulatory filings are rich with causal reasoning, yet extracting that reasoning automatically remains a difficult problem — one that sits at the intersection of domain expertise, linguistic nuance, and machine comprehension. In this paper, we describe our participation in the English subtask of the Financial Document Causality Detection shared task, FinCausal 2026, where systems are asked to identify verbatim causal spans from financial paragraphs in response to abstractive causal questions.Our approach is grounded in the intuition that a small, well-adapted model with the right inductive biases can outperform a larger but unfocused one. We fine-tune Qwen3-4B-Instruct-2507 on 2,000 domain-annotated instances using QLoRA, a parameter-efficient technique that enables meaningful adaptation under modest computational resources. Before training, we reformat all instances into the Qwen ChatML instruction template to align the model's generation behaviour with the verbatim extraction requirement of the task. At inference time, we further guide the model by retrieving the most causally relevant sentence from the context using TF-IDF cosine similarity, providing an explicit local signal before generation. Outputs are produced via greedy decoding to ensure deterministic, source-grounded predictions.Under the official LLM-as-a-judge evaluation framework — which scores responses on a 1–5 adequacy scale based on semantic correctness rather than lexical overlap — our system achieves a score of **4.76 out of 5**, placing **4th out of nine teams** on the English leaderboard. Our results suggest that combining instruction-tuned fine-tuning with lightweight retrieval is a practical and effective strategy for causal reasoning in specialised financial text.

Keywords: causal question answering, financial NLP, QLoRA, parameter-efficient fine-tuning, TF-IDF retrieval, span extraction, LLM-as-a-judge

1. Introduction

Financial documents such as earnings reports, annual filings, and regulatory disclosures do more than report numbers — they explain *why* those numbers changed. Understanding the causal reasoning embedded in these texts is fundamental to building systems that support financial analysis, risk assessment, and automated report generation. However, causality in financial language is rarely straightforward: causal relationships often span multiple sentences, involve compound contributing factors, and are expressed without explicit connectives such as *because* or *therefore* (Cheng et al., 2024; Girju, 2003).

The Financial Document Causality Detection shared task (FinCausal) (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023, 2025) directly addresses this challenge by evaluating systems on their ability to identify cause-and-effect relationships in financial text across English and Spanish. The 2026 edition introduces three significant advances over prior years: a revised dataset with richer and more complex causal annotations, the inclusion of multi-hop causal chains involving three or more events, and a new evaluation framework in which an LLM judge scores system responses on a 1–5 adequacy scale (Zheng et al., 2023), replacing the earlier exact-match and similarity-based metrics.

We participate in the English subtask and frame it as a **Causal Question Answering (CQA)** problem, where a system must extract the verbatim causal span from a financial paragraph in response to an abstractive question. Our approach consists of three components: (1) reformatting the training data into the Qwen ChatML instruction format, (2) fine-tuning Qwen3-4B-Instruct-2507 using QLoRA for parameter-efficient domain adaptation, and (3) an inference pipeline that combines intra-context TF-IDF retrieval with constrained greedy decoding to enforce verbatim extraction and reduce hallucination.

Our system achieves a score of **4.76 out of 5** on the official test set, demonstrating that a carefully fine-tuned compact language model, paired with a lightweight retrieval signal, can effectively identify causal relationships in complex financial text.

2. Related Work

2.1. Causality Detection in NLP

Early approaches to causality detection relied on lexical cues such as *because*, *therefore*, and *as a result* (Girju, 2003; Blanco et al., 2008), but

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 175–180

struggled with implicit causal expressions. Subsequent work introduced machine learning methods — SVMs, CNNs, and RNNs — that learned patterns directly from data (Do et al., 2011). More recently, transformer-based models such as BERT have become the dominant approach due to their contextual understanding (Cheng et al., 2024). Recent studies have also explored the use of prompt engineering with large language models to enhance causal relationship detection, particularly in domain-specific settings such as finance (Chatwal et al., 2025). Our work extends this line by fine-tuning the Qwen model on financial causal data.

2.2. Causal Question Answering

Extractive QA benchmarks such as SQuAD require answer spans to be identified directly within a context, but do not emphasise causal reasoning. Tasks like COPA (Gordon et al., 2012) target commonsense causality, while FinCausal (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023, 2025) focuses specifically on financial text. FinCausal 2026 raises the difficulty further by combining abstractive questions with extractive answers and introducing multi-hop causal chains, requiring models to reason rather than match.

2.3. Financial NLP

Financial text presents unique challenges for general NLP tools. Loughran and McDonald (2011) demonstrated that standard sentiment lexicons perform poorly on financial language, motivating domain-specific models such as FinBERT (Araci, 2019). Causality extraction in financial reports has gained traction through the FinCausal shared task series, which our work directly builds upon.

2.4. Retrieval-Augmented Generation and Parameter-Efficient Fine-Tuning

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020)(Didwania et al., 2024) has shown that grounding generative models with retrieved context improves factual accuracy. We adopt a lightweight variant of this idea, using TF-IDF cosine similarity to surface the most causally relevant sentences before generation. For efficient adaptation, we employ QLoRA (Dettmers et al., 2023), which enables fine-tuning of large language models under 4-bit quantization with minimal performance degradation. Finally, FinCausal 2026 adopts the LLM-as-a-judge evaluation framework (Zheng et al., 2023), which scores responses on semantic adequacy rather than lexical overlap — better reflecting the quality of causal reasoning.

3. Problem Statement

Financial narratives describe measurable changes in economic indicators and corporate performance, yet numerical values alone rarely explain *why* such changes occur. The Financial Document Causality Detection task, FinCausal 2026 (mor; Uniyal et al., 2021), addresses this gap by targeting **text-internal causal relationships** within financial documents. We formalize this as a **Causal Question Answering (CQA)** problem, where each instance is structured as a triplet (C, Q, A): a financial paragraph $C = \{w_1, w_2, \ldots, w_n\}$ serving as context, an abstractive causal question Q, and an extractive answer span $A \subseteq C$ drawn verbatim from the text.

A causal relation is defined as an ordered pair (e_c, e_e), where e_c is the causal event and e_e is the resulting event, such that $e_c \rightarrow e_e$. The task focuses strictly on how causality is *encoded within the document*, not on the real-world validity of the stated relationships.

3.1. Learning Objective

Given (C, Q), the goal is to learn a function $f_\theta : (C, Q) \rightarrow \hat{A}$, where $\hat{A} = C[i : j]$ is a contiguous extractive span. Model parameters are optimized by minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{N} \log P_\theta(A_k \mid C_k, Q_k) \tag{1}$$

Despite the extractive constraint, the task is posed in a **generative QA format** to handle complex multi-hop causal chains of the form $e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_k$, which are a key feature of the 2026 edition.

3.2. Evaluation

FinCausal 2026 replaces the earlier SAS + Exact Match scheme with an **LLM-as-a-judge** framework, where a judge model scores each response on a 1–5 adequacy scale. The overall system score is:

$$\text{Score} = \frac{1}{N} \sum_{k=1}^{N} J(A_k,\ \hat{A}_k) \tag{2}$$

This prioritizes **semantic adequacy and reasoning correctness** over strict lexical overlap.

4. Dataset

The FinCausal 2026 English dataset (Moreno-Sandoval et al., 2026) is sourced from UK financial annual reports (2017) compiled by UCREL at Lancaster University, supplemented with excerpts from the 2018 FinT-esp corpus. Relative to prior editions,

the dataset has been substantially revised: ambiguous and trivial instances were removed, and over 500 new fragments featuring complex causal structures — including chains of three or more events — were added. Approximately 10% of questions were rephrased to demand deeper reasoning beyond surface-level lexical matching. Training and test splits were constructed via random partitioning, ensuring uniform distribution of complex examples across both sets. We participated in the **English subtask**; Table 1 reports the split statistics.

Split	Language	Instances
Train	English	2,000
Test	English	500
Total		**2,500**

Table 1: Dataset statistics for the English subtask of FinCausal 2026.

4.1. Data Format

Each instance is a triplet (C, Q, A): a financial paragraph C as context, an abstractive causal question Q probing either the cause or the effect, and an extractive answer span A copied verbatim from C. The dataset is provided in CSV format with four fields: *ID*, *context*, *question*, and *answer* — where the answer field is withheld in the test set.

5. Methodology

Our approach to the FinCausal 2026 English subtask consists of two main stages: supervised fine-tuning of a compact instruction-tuned language model, and a retrieval-augmented inference pipeline designed to enforce verbatim causal span extraction. Algorithm 1 provides a high-level overview of the full system.

5.1. Base Model

We select **Qwen3-4B-Instruct-2507** (Team, 2025) as our base model. This model offers a strong balance between parameter efficiency and instruction-following capability, making it well-suited for a constrained extractive QA task on domain-specific financial text. Its compact size also allows fine-tuning and inference within limited computational budgets using quantization.

5.2. Input Formatting

Before fine-tuning, all training instances are reformatted into the **Qwen ChatML** template, which structures each example as a multi-turn conversation with explicit role markers. Each instance is formatted as follows:

Algorithm 1 FinCausal 2026 System Pipeline

Require: Context C, Question Q, Fine-tuned model f_θ
Ensure: Predicted causal span $\hat{A}$
// — Training Stage —
Reformat all (C, Q, A) instances into Qwen ChatML format
Load base Qwen3-4B-Instruct-2507 in 4-bit quantized mode
Attach QLoRA adapters to attention and feed-forward projections
Optimize θ by minimizing $\mathcal{L}(\theta) = -\sum_{k=1}^{N} \log P_\theta(A_k \mid C_k, Q_k)$
// — Inference Stage —
Load fine-tuned f_θ in 4-bit quantized mode
Tokenize C into sentences $\{s_1, s_2, \ldots, s_m\}$
Compute TF-IDF vectors for each s_i and Q
$s^* \leftarrow \arg\max_{s_i}$ cos(TF-IDF(s_i), TF-IDF(Q))
Construct prompt using full C, retrieved s^*, and Q
Generate $\hat{A} \leftarrow f_\theta(C, s^*, Q)$ using greedy decoding
return $\hat{A}$

Prompt Template

<|im_start|>system
You are a financial question answering assistant. Given a financial text, extract the answer to the causal question directly and **verbatim** from the context. Do not generate any answer that is not present in the text.

<|im_start|>user
Context: $[C_i]$
Question: $[Q_i]$

<|im_start|>assistant
$[A_i]$
<|im_end|>

This formatting aligns the training distribution with the model's pre-trained instruction-following behaviour, ensuring that the model learns to respond within the expected conversational structure rather than treating the task as raw text completion.

5.3. QLoRA Fine-Tuning

We fine-tune the model using **QLoRA** (Quantized Low-Rank Adaptation) (Dettmers et al., 2023), which combines 4-bit quantization of the base model weights with low-rank adapter layers inserted into the transformer architecture. This significantly reduces GPU memory requirements while preserv-

ing the model's ability to adapt to the target domain.

Low-rank adapters are injected into all major projection layers of the transformer, including the attention projections ($\mathbf{q}$, $\mathbf{k}$, $\mathbf{v}$, $\mathbf{o}$) and the feed-forward projections (`gate_proj`, `up_proj`, `down_proj`). Targeting the feed-forward layers in addition to the attention layers has been shown to substantially improve task-specific adaptation (Dettmers et al., 2023). Table 2 summarises the LoRA configuration used.

Hyperparameter	Value
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.05
Bias	none
Task type	Causal LM
Target modules	q, k, v, o, gate, up, down proj
Quantization	4-bit (NF4)

Table 2: QLoRA fine-tuning configuration.

The rank $r = 32$ provides sufficient adapter capacity for capturing financial domain patterns, while $\alpha = 64$ (set to $2r$ following standard practice) controls the scaling of the adapter updates. A dropout of 0.05 is applied to the adapter layers as a lightweight regularisation measure.

5.4. Inference Pipeline

At inference time, we apply a four-step pipeline designed to minimise hallucination and enforce verbatim extraction from the source context.

Step 1 — Model Loading. The fine-tuned model is loaded in **4-bit quantized mode** using BitsAndBytes, reducing memory footprint while maintaining generation quality.

Step 2 — Intra-Context Retrieval. Rather than passing the full context blindly to the model, we apply an **intra-context retrieval** step. Each sentence in the context C is ranked against the question Q using **TF-IDF cosine similarity**. The top-ranked sentence s^* is selected as the most relevant causal evidence:

$$s^* = \arg\max_{s_i \in C} \cos(\text{TF-IDF}(s_i),\ \text{TF-IDF}(Q)) \quad (3)$$

This retrieved sentence is appended to the prompt alongside the full context, providing the model with an explicit signal about where the causal span is likely to reside.

Step 3 — Constrained Prompt. The model is prompted using a **verbatim extraction format** that explicitly instructs the model to copy the answer word-for-word from the context. Both the full context and the retrieved relevant sentence are included in the prompt, reducing the risk of paraphrased or hallucinated responses.

Step 4 — Greedy Decoding. Generation is performed using **greedy decoding** with temperature set to zero, ensuring fully deterministic outputs. This choice is deliberate: since the task requires precise verbatim spans, stochastic sampling strategies such as top-p or beam search with diversity penalties are counterproductive. Greedy decoding directly maximises the probability of the most likely token at each step, producing stable and reproducible predictions aligned with the source text.

6. Results

6.1. Main Result

We evaluate our system on the FinCausal 2026 English test set comprising 500 instances, using the official LLM-as-a-judge metric. Our system achieves a score of **4.76 out of 5**, ranking **4th** on the official English subtask leaderboard out of nine participating teams. Unlike Exact Match, the LLM judge tolerates minor boundary differences provided the causal meaning is preserved, making it a more faithful measure of reasoning quality in financial text. Table 3 reports the full leaderboard standings.

Rank	Team	LLM Score
1	HSA_CORAL	4.814
1	Sheffield_Causal	4.814
3	Tredence_AICOE	4.812
4	**Lab Rats (Ours)**	**4.760**
5	CariMed	4.720
6	EMI	4.704
7	LeedsMeng26	4.700
8	TU Graz Data Team	4.662
9	Sarang	4.540

Table 3: Official English subtask leaderboard for FinCausal 2026.

6.2. Component Analysis

Three design choices contribute to the strong performance of our system.

QLoRA Fine-Tuning. Training on 2,000 expert-annotated instances adapts the Qwen model to the

vocabulary, syntax, and causal patterns of financial reporting. Without this step, the base model lacks the domain grounding needed to distinguish causal spans from surrounding narrative text.

ChatML Instruction Formatting. Structuring each training instance as a ChatML instruction explicitly conditions the model to extract answers verbatim from the context rather than paraphrase or hallucinate plausible-sounding responses.

TF-IDF Intra-Context Retrieval. Ranking sentences by cosine similarity to the question before generation provides the model with an explicit relevance signal. This is particularly effective for instances where the question and causal span share limited lexical overlap, a common characteristic of abstractive causal questions in the 2026 dataset.

7. Limitations

Despite strong overall performance, our system faces two notable limitations. First, predicting long answer spans — particularly those covering multiple clauses — remains challenging, as small boundary errors can reduce semantic fidelity. Second, multi-hop causal chains involving three or more events are difficult to resolve through span extraction alone, as they require discourse-level reasoning that goes beyond identifying a single contiguous passage. Addressing these limitations likely requires models with explicit coreference and discourse structure awareness, rather than relying solely on local retrieval signals.

8. Conclusion

We presented a system for the FinCausal 2026 English subtask that combines parameter-efficient fine-tuning with a lightweight retrieval-augmented inference pipeline. By reformatting training data into the Qwen ChatML instruction format, fine-tuning Qwen3-4B-Instruct-2507 via QLoRA, and augmenting inference with TF-IDF intra-context retrieval and greedy decoding, our system achieves a score of **4.76 out of 5** on the official evaluation.

Our results demonstrate three broader findings. First, compact language models fine-tuned on modest domain-specific datasets can achieve strong performance on financial causal QA when paired with appropriate instruction formatting. Second, lightweight retrieval signals such as TF-IDF remain effective for grounding generation even without dense retrieval infrastructure. Third, the LLM-as-a-judge evaluation framework provides a more meaningful signal than lexical overlap metrics for tasks requiring causal reasoning, rewarding semantic correctness over verbatim matching.

Future work should explore discourse-aware models capable of resolving multi-hop causal chains, as well as dense retrieval methods that better capture semantic similarity between abstractive questions and their corresponding causal spans in financial text.

9. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and paraphrasing during the preparation of this manuscript, including grammar correction and improving the clarity of written expressions. No generative AI tool was used to produce any scientific content, experimental results, analysis, or conclusions presented in this work. All authors are fully responsible and accountable for the content of this paper.

10. Acknowledgements

This research was conducted independently by the authors outside the scope of their professional responsibilities at Sony Research. The views and findings presented in this paper are solely those of the authors and do not reflect the positions or policies of Sony Research.

11. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Pulkit Chatwal, Amit Agarwal, and Ankush Mittal. 2025. Enhancing causal relationship detection using prompt engineering and large language models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 248–252.

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2024. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *arXiv preprint arXiv:2411.10371*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Krish Didwania, Pratinav Seth, Aditya Kasliwal, and Amit Agarwal. 2024. Agrillm: harnessing transformers for framer queries. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 179–187.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 76–83.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 105–107.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (fincausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (fincausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

Qwen Team. 2025. Qwen3 technical report.

Deepak Uniyal, Amit Agarwal, Durga Toshniwal, and Dipanjan Deb. 2021. Dense vector embedding based approach to identify prominent disseminators from twitter data amid covid-19 outbreak. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(3):308–320.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Author Index