LREC 2026

Learning Non-Literal Expressions with Small Data @ LREC 2026

Workshop Proceedings

Editors
Markus Egg and Valia Kordoni

11 May 2026

ISBN 978-2-493814-80-7

Preface

Non-Literal Expressions (NLEs) in natural language are a reflection of fundamental cognitive processes such as analogical reasoning and categorisation, and are deeply rooted in everyday communication. NLEs understanding is therefore an essential task for language modeling. This task is especially challenging because it cannot be tackled by falling back on individual word meanings, but requires taking into account larger chunks of surrounding text or even contextual information. At the same time, it is important because the reliable processing of NLEs is relevant for optimizing downstream tasks like translation and summarization.

This workshop focuses on understanding of Non-Literal Expressions. While most of the earlier work on NLEs had been devoted to metaphor and metonymy, recent activities target other forms of NLEs as well, e.g., hyperbole (deliberate exaggeration), litotes (understatement), rhetorical questions, and irony. Humanly annotated corpora for NLEs have very recently started becoming available to the research community and may serve as the basis for data-driven approaches to NLEs processing, with the interrelated goals of first identifying and then interpreting such expressions. Such data is mostly of high linguistic quality, but still very limited in size. Thus, the workshop's focus is on adaptation of Language Models (LMs) and Deep Learning (DL) for processing of Non-Literal Expressions with limited high-quality data, since such constructs still pose big identification and processing challenges in natural language analysis tasks.

The workshop features contributions which focus on the use of techniques like self-training for leveraging unlabelled data, as well as in work that focuses on the incorporation of external linguistic resources and knowledge injection to enrich features, and also in research that describes work on utilisation of multitask learning with the aim to benefit from related tasks.

The workshop also wants to discuss alternative approaches which may elaborate on the use of pre-trained Language Models (LMs) as a foundation and the application of techniques like contrastive learning and clustering to identify challenging examples within the data, the ultimate aim of the workshop being to highlight the necessity of high-quality data, as well as cross-lingual datasets.

Organizing Committee

Markus Egg, Humboldt-Universität zu Berlin, Germany
Valia Kordoni, Humboldt-Universität zu Berlin, Germany

Programme Committee

Beata Beigman Klebanov, ETS, USA
Maria Berger, Ruhr-Universität Bochum, Germany
Yuri Bizzoni, Aarhus University, Denmark
Kenneth Church, VecML Inc., USA
Stefanie Dipper, Ruhr-Universität Bochum, Germany
Markus Egg, Humboldt-Universität zu Berlin, Germany
Anna Feldman, Montclair State University, USA
Debanjan Ghosh, Princeton, USA
Valia Kordoni, Humboldt-Universität zu Berlin, Germany
Emmy Liu, CMU, USA
Petya Osenova, Sofia University "St. Kl. Ohridski", Bulgaria
Sebastian Padó, IMS Stuttgart, Germany
Gudrun Reijnierse, Vrije Universiteit Amsterdam, The Netherlands
Sebastian Reimann, Ruhr-Universität Bochum, Germany
Adam Roussel, Ruhr-Universität Bochum, Germany
Tatjana Scheffler, Ruhr-Universität Bochum, Germany
Sabine Schulte im Walde, Universität Stuttgart
Vered Shwartz, The University of British Columbia, Canada
Caroline Sporleder, Georg-August-Universität Göttingen, Germany
Egon Stemle, EURAC, Italy

Table of Contents

Workshop Program

Monday, May 11, 2026

9:00–13:00 **Learning Non-Literal Expressions with Small Data**
Room: 4
Chair: Valia Kordoni

9:00–9:10 ***Introduction***

9:10–9:50 *Challenges in Japanese Euphemism Classification: An Analysis of Pre-trained Japanese and Multilingual Models*
Noriko Takahashi, Whitney Poh, Libby Barak, JIng Peng and Anna Feldman

9:50–10:10 *Steering Pragmatic Interpretation in LLMs: A Diagnostic Evaluation of Few-Shot and Reasoning-Based Prompting for Indirect Speech Acts.*
Massimiliano Orsini and Dominique Brunato

10:10–10:30 *Injecting Structured Lexicographic Knowledge into LLMs for Non-Literal Expression Disambiguation: A Controlled Study on Croatian*
Slobodan Beliga, Ivana Filipović Petrović and Ana Meštrović

10:30–11:00 ***Coffee break***

11:00–11:40 **Poster session**

Metaphor Identification in Spanish Oncological Discourse: The Role of Explicit Meaning in Low-Resource Settings
Lucia Pitarch, Jordi Bernad and Gemma Bel-Enguix

Exploring Detection of Complex, Non-Literal Expressions of Cultural Motifs
Ibrahim H. Alyami and Mark A. Finlayson

Artful Writing, Authentic Emotions: Distinguishing Human-Written from LLM-Generated Metaphors by Annotation and Classification
Michaela Regneri, Nooshin Aghajari and Thomas Kroedel

Creation and Validation of a Monolingual Spanish NLI Dataset for Metaphor Interpretation via Model-in-the-Loop
Alec Sanchez-Montero, Gemma Bel-Enguix and SERGIO LUIS OJEDA TRUEBA

A Hybrid Architecture for Metonymy Detection in Marathi
Pratibha Dongare

Monday, May 11, 2026 (continued)

Contextualising (Im)plausible Events Triggers Figurative Language
Annerose Eichel, Tonmoy Rakshit and Sabine Schulte im Walde

11:40–12:00 *A Novel Dataset and Three Ways to Approach Automatic Metaphor Detection in German Religious Online Forums*
Sebastian Reimann and Tatjana Scheffler

12:00–12:20 *Decomposing Creativity: Two Small Datasets Combining Originality Ratings and Metaphor Annotations*
Emilie Sitter, Sina Zarrieß, Omar Momen and Berenike Herrmann

12:20–13:00 *Unveiling Reasoning in Small Language Models: Insights into Literal and Non-Literal Understanding*
Debanjan Gosh

Challenges in Japanese Euphemism Classification: An Analysis of Pretrained Japanese and Multilingual Models

Noriko Takahashi, Whitney Poh, Libby Barak, Jing Peng, Anna Feldman
Montclair State University
1 Normal Ave, Montclair, NJ 07043, USA
{takahashin1, pohw1, barakl, pengj, feldmana}@montclair.edu

Abstract

Euphemisms present a persistent challenge for NLP because their interpretation depends on pragmatic inference, social norms, and contextual cues rather than surface meaning alone. Although Potentially Euphemistic Terms (PET)-based resources have been developed for several languages, Japanese euphemisms remain computationally unexplored despite their close interaction with honorifics, register variation, and orthographic choice. We introduce **JP-PET**, the first PET-based dataset for Japanese euphemism classification, comprising 1,672 annotated sentences across 101 PETs and ten semantic domains with register metadata. We evaluate two Japanese monolingual transformer models (Rinna RoBERTa and Tohoku BERT) and the multilingual XLM-R under three controlled PET-level data splits that isolate lexical familiarity and generalization to unseen euphemisms. While models achieve strong performance when PETs are shared between training and test data, performance drops substantially under PET-disjoint conditions, indicating reliance on lexical familiarity. Error analysis suggests potential challenges in politically conventionalized expressions, metaphor-based euphemisms, and orthographic mitigation strategies. JP-PET provides the first benchmark for studying pragmatic meaning in Japanese NLP.

Keywords: Japanese NLP, euphemism detection, pragmatic inference, PETs, language models, register variation

1. Introduction

Natural Language Processing (NLP) continues to face challenges in interpreting indirect language uses, where meaning depends on pragmatic inference and cultural knowledge rather than literal expression. Recent work shows that despite substantial progress on various NLP tasks, pretrained transformer still struggle with the processing non-literal expressions, including figurative language that is novel in the sense of being rare, recently coined, or not frequently observed during pretraining, as well as culturally specific expressions (Liu et al., 2022; Jang et al., 2023; Ichien et al., 2024). Euphemisms pose a related challenge: they soften or reframe direct meanings to maintain politeness or avoid discomfort, so interpretation depends on social norms and contextual cues rather than surface semantics (Allan and Burridge, 1991; Pinker, 2003). Prior studies show that pretrained transformer models often struggle with euphemisms, whose interpretation depends on context, pragmatic inference, or distinguishing literal from euphemistic uses, as well as expressions involving metonymy and metaphor (Gavidia et al., 2022; Lee et al., 2024). These limitations highlight a central challenge for euphemism research: distinguishing between the literal and euphemistic uses of Potentially Euphemistic Terms (PETs). Many PETs retain their literal meaning in some contexts while becoming euphemistic in others, making them difficult for models to classify reliably. For example, the phrase “the bird and the bees” could refer to actual birds and bees as animals, but also refer to the sex-ed talk.

While recent work in English (Lee et al., 2024; Gavidia et al., 2022; Lee et al., 2023) has begun to address this challenge by developing PET frameworks and euphemism-detection benchmarks, these efforts remain almost entirely language-specific. Japanese euphemisms, in particular, remain largely unexplored, even though Japanese communication is strongly shaped by indirectness, contextual interpretation, and sensitivity to social relationships (Maynard, 1997; Cook, 2006). As pretrained transformer continue to be integrated into education, translation, writing assistance, and customer-facing computer applications, it is crucial to evaluate how well they handle these nuanced forms of indirect meaning. Understanding pretrained transformer performance in regards to Japanese euphemisms will not only highlight current limitations but will also inform the development of systems that must operate reliably in culturally sensitive contexts. This study, therefore, examines Japanese euphemistic expressions and evaluates how pretrained transformer interpret them, with the goal of supporting future applications that require accurate and contextually appropriate handling of Japanese pragmatic meaning.

To address the gap identified above, and to establish a foundation for computational work on Japanese euphemisms, this study makes three primary contributions.

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 1–11
11 May 2026.

- Introduces the first PET-based Japanese euphemism dataset, including metadata such as domain, register, and orthographic variation.
- Evaluates Japanese and multilingual pretrained models on distinguishing literal vs. euphemistic PET usage.
- Conducts a focused error analysis to identify key challenges in modeling Japanese euphemisms.

2. Potentially Euphemistic Terms (PETs)

To capture the ambiguity inherent in euphemistic expressions, NLP researchers introduced the concept of potentially euphemistic terms (PETs), which are words or phrases whose meaning can be literal or euphemistic depending on context (Lee et al., 2022; Gavidia et al., 2022). For instance, *between jobs* may literally describe a temporary career transition (e.g., *She is between jobs after finishing her contract*) or function euphemistically to mean unemployed (e.g., *He has been between jobs for over a year*) (Lee et al., 2022). Similarly, *special* may denote a general sense of uniqueness but can also serve as a euphemism for disability in expressions such as *special needs*. These examples illustrate how a single expression can function as a euphemism in some contexts but not in others. Because language and social norms evolve, and because interpretation varies across individuals, annotators often disagree on whether an expression should be labeled as euphemistic. PETs therefore offer a practical framework for annotation and modeling. As noted by Gavidia et al. (2022), corpora annotated for PETs capture this pragmatic variability and highlight the importance of context-sensitive interpretation. Modeling PETs requires distinguishing pragmatic meaning from surface semantics, which makes the task considerably more complex than standard lexical classification.

3. Linguistic Characteristics of Euphemisms

Euphemisms are expressions that allow speakers to avoid direct meaning, hide uncomfortable truths, or convey politeness through indirect speech in order to maintain social relationships (Rababah, 2014). As Allan and Burridge (1991) explain, euphemistic expressions soften the impact of a message and avoid taboo wording, and they commonly appear in culturally sensitive domains such as death, illness, sexuality, disability, aging, and social class (Casas Gómez, 2009; Valentine, 1998). Because euphemisms are tied to social norms, their meanings vary across communities and registers and change over time. Pinker (2003) describes this process as the ”euphemism treadmill,” in which expressions gradually lose their softening effect as they become associated with the taboo concepts they refer to, leading speakers to introduce new forms (Pinker, 2007).

Euphemisms are often realized through strategies such as circumlocution, lexical substitution, abbreviation, and semantic generalization (Allan and Burridge, 2006). While these mechanisms are broadly shared across languages, the specific expressions and their social motivations are language-specific, which makes their computational modeling highly dependent on cultural and contextual knowledge.

4. Euphemisms in Japanese

4.1. Domains of Japanese Euphemisms

Japanese euphemisms appear across common taboo domains such as *Death & Dying, Bodily Functions, Sexuality & Relationships, Crime & Social Issues,* and *Illness & Disability* (Allan and Burridge, 1991; Casas Gómez, 2009). They are also frequently used in domains such as Appearance, Personality, *Workplace & Economy*, and *Aging*. For example, 旅立つ (tabidatsu; literal meaning: to set out on a journey) is used euphemistically to mean ‘to die’ (Maynard, 1997), and ぽっちゃり (pochari) provides a Japanese mimetic expression that conveys a softened description of body size (Hamano, 1998).

Although these domains broadly overlap with those observed in English, the specific expressions and their sociocultural motivations differ, reflecting language-specific norms of politeness and indirectness.

4.2. Japanese Orthography

Japanese orthography allows a single lexical item to be written in kanji, hiragana, or katakana. These choices can signal differences in tone and social nuance. Orthographic variation therefore plays an important role in euphemistic expression by softening negative connotations or creating distance from sensitive meanings. For example, the word 障害 (shōgai) ‘disability’ is sometimes written as 障がい, where the kanji 害 (‘harm’) is replaced with the hiragana がい to reduce harshness. Another example involves katakana: the kanji 物 (mono/butsu) ‘thing’ can appear as ブツ (butsu), a slang form that can refer to drugs or contraband and thus carries a more implicit, coded meaning. By using katakana instead of kanji, the term de-

parts from the literal sense ‘thing’ and signals an indirect reference.

These variations introduce additional challenges for computational modeling, as identical pronunciations may correspond to different pragmatic functions depending on the script. Models must therefore recognize orthographic choices as signals of pragmatic intent rather than treating all forms as equivalent.

4.3. Register Variation

Register	PET	Literal	Meaning
Formal	お亡くなりになる (onakunari ni naru)	gone	to pass away
Neutral	女の子の日 (onnanoko no hi)	girl’s day	menstruation
Informal	ボンビー (bonbii)	–	‘poor’

Table 1: Examples of Japanese euphemistic expressions (PETs) across registers, with literal gloss and euphemistic meaning.

Register refers to the level of politeness and social distance conveyed through linguistic expression, and in Japanese, euphemisms often vary by register even when they express the same underlying meaning (Maynard, 1997). In this study, we group Japanese euphemisms into three levels: formal, neutral, and informal, reflecting differences in situational context and social expectations (Biber and Conrad, 2019). These variations create challenges for computational models, as the same meaning can be realized through forms that differ in politeness and social intent. Models must therefore rely on contextual cues beyond surface meaning to correctly interpret euphemistic usage. See Table 1 for an example for each level. The informal example ボンビー (bonbii) does not have a direct literal meaning, as it is a phonological modification of 貧乏 (binbo-) ’poor’ used to soften or stylize the expression.

5. Research Questions

Japanese euphemisms challenge computational modeling because they vary across domains, permit orthographic alternations, and shift register with social context. Because PET-based resources do not capture these Japanese-specific patterns, we establish a baseline for Japanese euphemism classification using pretrained Japanese and multilingual models and analyze their generalization across controlled PET-level splits.

RQ1. How reliably do pretrained Japanese and multilingual language models classify euphemistic expressions across domains and registers?

RQ2. In which domains and registers do these models most frequently produce errors, and what insights do these patterns provide for future euphemism modeling and dataset construction?

Here, domains refer to semantic fields, and register refers to politeness level (formal, neutral, informal).

6. Methodology

This study frames euphemism detection as a binary classification task. For each sentence containing a potentially euphemistic term (PET), we insert [PET_BOUNDARY] markers around the target expression and provide the sentence to the model as input. The model predicts whether the marked PET is used euphemistically (1) or literally (0). We fine-tune three pretrained transformer models Rinna (rinna Co., Ltd., 2023), Tohoku (University, 2020), and XLM-Roberta (Conneau et al., 2020), on this task and evaluate their generalization across two PET-level data splits designed to test familiar-item learning and strict no-overlap generalization to new euphemisms.

6.1. PETs List and Data Collection

We collected data from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) via Shonagon, which covers diverse genres (e.g., news, magazines, books, web, official documents) and supports variation in register and orthography. We compiled 100 PETs based on prior PET work (Lee et al., 2022; Biyik et al., 2024), Japanese pragmatics and politeness research (Maynard, 1997; Cook, 2006), and native-speaker judgment, and queried BCCWJ to retrieve naturally occurring sentences containing inflected and orthographic variants.

6.2. Automatic PET Extraction

We used Sudachi (Takaoka et al., 2018) for morphological analysis and customized rules to capture inflectional and orthographic variants of PETs. Detected PET spans were marked with [PET_BOUNDARY] so the model could focus on the target region (e.g., Input: 「彼は去年、亡くなりました。」’He passed away last year.’ Annotated:「彼は去年、[PET_BOUNDARY] 亡くなりました [PET_BOUNDARY]。」‘He [PET_BOUNDARY] passed away [PET_BOUNDARY] last year.’).

6.3. Annotation Procedure

Three native Japanese speakers annotated each PET instance as euphemistic (1) or literal (0) us-

Model	Type	Tokenization	Pretraining Data	Coverage	Main Characteristics
Rinna	Monolingual	SentencePiece (subword)	Japanese CC-100 + Wikipedia	Japanese	Broad web coverage; captures modern usage and stylistic variation
Tohoku BERT	Monolingual	MeCab + UniDic → WordPiece	Japanese Wikipedia	Japanese	Linguistically informed segmentation; encyclopedia-style language
XLM-R	Multilingual	SentencePiece (subword)	Multilingual CC-100	100+ languages	Large-scale multilingual training; strong generalization

Table 2: Comparison of models, tokenization, and pretraining data.

ing shared guidelines with examples of clear and ambiguous cases. Inter-annotator agreement was high (Fleiss' κ = 0.87, 95% CI [0.85, 0.89]), and final labels were determined by majority vote.

6.4. Dataset Overview

JP-PET contains 1,672 annotated sentences for 101 PETs spanning multiple domains (Table 3) Per-PET statistics are reported in Appendix A.1.

Domain	Count	Percent (%)
Death & Dying	406	24.3
Bodily Functions	299	17.9
Sexuality & Relationships	224	13.4
Crime & Social Issues	168	10.1
Illness & Disability	146	8.7
Appearance	98	5.9
Personality	86	5.1
Workplace & Economy	83	5.0
Aging	69	4.1
Others	63	3.8
Politics & War	30	1.8
Total	1,672	100.0

Table 3: Distribution of domains in the JP-PET dataset.

Note: Domains correspond to semantic categories commonly observed in euphemism research (Allan and Burridge, 2006).

6.5. Dataset Splitting

To investigate how Japanese euphemisms vary in difficulty across semantic domains and individual Potentially Euphemistic Terms (PETs), we designed two complementary data-splitting schemes. Both splits use the same canonical test pool to ensure that model performance is directly comparable across conditions. In Split 1, PETs may appear in the training, validation, and test sets, allowing lexical overlap. In Split 2, PETs in the test set are strictly unseen during training, enforcing a no-overlap condition.

6.5.1. Always vs. sometimes euphemisms

For each PET, we first examined its label distribution. PETs annotated as euphemistic in all occurrences (label = 1) were categorized as always-euphemistic, whereas PETs that appeared with both euphemistic and non-euphemistic labels were categorized as sometimes-euphemistic. Because always-euphemistic expressions provide stable pragmatic meaning, all of their sentences were reserved for the test pool to probe model performance on consistently euphemistic items.

6.5.2. Canonical shared test pool and 10 folds

The canonical test pool consisted of:

- all sentences containing always-euphemistic PETs, and
- a stratified sample of sentences for sometimes-euphemistic PETs, balanced across labels (1 = euphemistic, 0 = non-euphemistic) for each PET.

This pool was partitioned into 10 disjoint folds, ensuring that across the 10 folds, every PET instance appears exactly once in a test set. This design enables fine-grained inspection of model performance on all PETs while keeping the evaluation consistent across both splits.

6.5.3. Split 1 – Shared PETs across train/validation/test

In Split 1, PETs are allowed to appear in all partitions. After constructing the canonical test pool, the remaining sometimes-euphemistic sentences were divided into training and validation sets using label-balanced sampling within each PET. The test sets correspond exactly to the predefined canonical folds; no additional test instances were created.

The test folds (Split 1-1 through 1-10) follow the fixed test partitions, while train and validation remain constant across folds. This split reflects a standard setting in which the model encounters the same PET string during training and testing but in

different contexts, allowing us to measure contextual generalization without lexical novelty.

6.5.4. Split 2 – PET-disjoint train/validation/test

Split 2 reuses the exact same 10 test folds as Split 1 but enforces complete PET disjointness across all partitions. No PET in the test set appears in train or validation, and no PET in validation appears in train. For each fold, entire PETs (not single sentences) were assigned to train or validation, while keeping label balance within each PET as much as possible. This split evaluates zero-shot PET generalization – whether a model can recognize euphemistic meaning for lexically unseen PETs.

6.5.5. Data Partitioning and Label Balancing

Across all splits, data are partitioned into training, validation, and test sets using an approximate 8:1:1 ratio, subject to PET-level constraints. For sometimes-euphemistic PETs, we perform label balancing by sampling approximately equal numbers of euphemistic (1) and non-euphemistic (0) instances within each PET and split. Always-euphemistic PETs appear exclusively in the test sets, which facilitates comparison of model behavior for stable euphemisms across Splits 1–2.

The distribution of test instances by domain and label is reported in Appendix A.2. Within the sometimes-euphemistic subset, the test data remain close to balanced across labels. Small imbalances occur for PETs with inherently uneven label distributions, where perfect balancing is not possible. Nevertheless, the overall distribution remains close to balanced and supports stable evaluation across domains.

The dataset is publicly available.[1]

6.6. Models

We fine-tuned three transformer models (Table 2): Rinna (monolingual RoBERTa; SentencePiece; (rinna Co., Ltd., 2023)), Tohoku (monolingual BERT; MeCab/UniDic + WordPiece; (University, 2020)), and XLM-R (multilingual; SentencePiece; (Conneau et al., 2020)). We used max length 320, batch size 16, learning rate 2×10^{-5}, up to 8 epochs with early stopping, and selected models by validation macro-F1. Results are averaged over five runs.

As a baseline, we train a linear classifier over a frozen encoder. Each input instance is encoded by a pretrained model, and the final-layer [CLS] representation is used as a fixed feature vector. A logistic regression classifier is trained on these [CLS] representations without updating the encoder. The baseline is trained and evaluated on the same data splits as the fine-tuned models.

[1] https://github.com/NLPlabMSU/jp-pet-dataset.git

	Split 1 (Seen)		Split 2 (Unseen)	
Model	Baseline	Fine-tuned	Baseline	Fine-tuned
Rinna RoBERTa	0.45	0.68	0.46	0.57
Tohoku BERT	0.50	**0.73**	0.51	0.59
XLM-RoBERTa	0.48	**0.73**	0.52	**0.61**

Table 4: Macro-F1 scores for baseline and fine-tuned models across Split 1 (seen expressions) and Split 2 (unseen expressions). The baseline uses a frozen encoder with a linear classifier.

6.7. Evaluation Metrics

We report macro-F1 and per-label F1 ($F1_1$, $F1_0$). Thresholds are tuned on development data, and final results are averaged across runs. We additionally compute domain- and register-level error rates for error analysis.

7. Results

7.1. RQ1. How reliably do the models classify euphemisms?

Table 4 presents macro-F1 scores for baseline and fine-tuned models across Split 1 (seen expressions) and Split 2 (unseen expressions).

All models achieve their highest performance in Split 1, where the training, validation, and test sets share euphemistic expressions. In this setting, fine-tuning substantially improves performance for all models. XLM-R and Tohoku BERT achieve the highest macro-F1 (0.73), with large gains over their baselines (from 0.48 to 0.73 and from 0.50 to 0.73, respectively), while Rinna shows a smaller improvement (from 0.45 to 0.68). These results suggest that models benefit from lexical overlap and can learn expression specific patterns when the same euphemisms appear in both training and test data.

Performance drops in Split 2, where all test expressions are unseen. Macro-F1 falls to 0.57–0.61 across models, indicating that generalization to new euphemistic expressions remains challenging. Although fine-tuning still improves performance (e.g., XLM-R from 0.52 to 0.61, Rinna from 0.46 to 0.57), the gains are smaller than in Split 1. This gap suggests that a substantial portion of model performance is driven by lexical familiarity, and that models struggle to interpret euphemistic meaning when surface forms differ.

There are also differences across models. XLM-R shows the largest improvement in Split 1 (+0.25), followed by Tohoku BERT (+0.23) and Rinna (+0.23). In Split 2, XLM-R achieves the highest macro-F1 (0.61), indicating relatively stronger generalization to unseen expressions. Tohoku performs comparably (0.59), while Rinna shows lower performance (0.57), suggesting weaker generalization to new euphemistic expressions.

We observe an apparent per-label performance gap when evaluating on the full canonical test pool, where F1 is consistently higher for euphemistic instances (label 1; F1 ≈ 0.83–0.89) than for literal instances (label 0; F1 ≈ 0.29–0.57) (see Appendix A.3). This gap is partly influenced by test-set composition, as the canonical pool includes always-euphemistic PETs that are labeled only as euphemistic.

To examine this bias, we additionally evaluate on sometimes-euphemistic PETs only, which are label-balanced within each split. Under lexical overlap (Split 1), models achieve relatively balanced performance across labels (e.g., $F1_1$ ≈ 0.75–0.84, $F1_0$ ≈ 0.62–0.80). However, under no-overlap generalization (Split 2), F1 for label 0 drops substantially (≈0.40–0.56), while F1 for label 1 remains higher (≈0.63–0.68). This persistent gap suggests that models tend to over-predict euphemistic usage and rely on lexical associations.

On always-euphemistic PETs, models achieve high F1 for label 1 (≈0.88–0.95), indicating strong performance on lexically stable euphemisms.

Overall, the results suggest that current models capture useful cues for euphemistic usage but rely heavily on lexical familiarity and tend to over-predict positive cases. This highlights the importance of evaluating per-label behavior when assessing pragmatic generalization.

7.2. RQ2. Where do the models produce errors, and why?

To examine model behavior beyond aggregate performance, we conducted error analysis by semantic domain and register. For each domain and register category, we computed the error rate as the proportion of incorrectly classified instances among all test instances in that category:

$$\text{Error Rate} = \frac{\#\text{Incorrect}}{\#\text{Total Instances}}$$

Error rates were calculated separately for each model and split. Because these values represent proportions of misclassified instances, higher values indicate poorer performance within a given category.

Domain	Rinna			Tohoku			XLM-R		
	Base	S1	S2	Base	S1	S2	Base	S1	S2
Aging (46)	57%	13%	27%	38%	13%	31%	54%	8%	19%
Appearance (51)	42%	25%	18%	44%	37%	31%	70%	39%	47%
Bodily Functions (134)	47%	23%	31%	49%	20%	31%	39%	20%	30%
Crime & Social Issues (47)	47%	21%	38%	46%	25%	23%	28%	19%	27%
Death & Dying (208)	51%	10%	16%	35%	7%	18%	40%	9%	16%
Illness & Disability (29)	44%	29%	**39%**	55%	19%	**55%**	29%	19%	35%
Others (28)	40%	11%	14%	53%	7%	11%	60%	14%	7%
Personality (48)	59%	19%	19%	54%	13%	27%	69%	13%	19%
Politics & War (30)	**63%**	**50%**	23%	**78%**	**63%**	50%	**95%**	**67%**	**53%**
Sexuality & Relationships (113)	51%	15%	20%	32%	12%	19%	31%	8%	20%
Workplace & Economy (43)	61%	24%	33%	63%	22%	38%	63%	24%	38%

Table 5: Domain-level **Error Rates** (%) for baseline (Base), Split 1 (S1: seen expressions), and Split 2 (S2: unseen expressions). Lower values indicate better performance and boldface indicates the highest error rate within each column (i.e., vertical comparison).

7.2.1. Per Domain

Table 5 reports domain-level error rates for Splits 1–2. We analyze each split separately to identify which semantic domains are most challenging.

Across models and splits, *Death & Dying* consistently shows the lowest error rates (e.g., 7–10% in Split 1 and around 16–18% in Split 2), indicating that this domain is the most reliably classified. This suggests that euphemisms related to death are highly conventionalized and semantically transparent, making them easier for models to recognize. This domain also contains the largest number of test instances (n=208), which contributes to the stability of the observed error rates.

In contrast, *Politics & War* exhibits the highest error rates across models (e.g., up to 67% in Split 1 and above 50% in Split 2). Errors remain elevated in both splits, indicating that this domain is consistently difficult regardless of lexical overlap. This pattern is largely driven by expressions such as 遺憾 (*ikan*) 'regrettable'. Across models, 遺憾 shows a substantially higher error rate than other items in this domain and accounts for a large proportion of domain-level errors. Removing this expression leads to a noticeable reduction in overall error, suggesting that a small number of politically conventionalized euphemisms disproportionately affect model performance.

The *Illness & Disability* domain also remains challenging, particularly under unseen conditions. Error rates increase substantially in Split 2 for all models (e.g., from around 19% to over 50% for

Register	Rinna			Tohoku			XLM-R		
	Base	**S1**	**S2**	**Base**	**S1**	**S2**	**Base**	**S1**	**S2**
Formal (89)	**67%**	17%	17%	**51%**	**24%**	**29%**	49%	**26%**	22%
Informal (82)	27%	13%	18%	36%	10%	18%	28%	12%	24%
Neutral (623)	51%	**19%**	**25%**	45%	17%	27%	**50%**	17%	**26%**

Table 6: Register-level error Rates (%) for baseline (Base), Split 1 (S1: seen expressions), and Split 2 (S2: unseen expressions). Lower values indicate better performance and boldface indicates the highest error rate within each column (i.e., vertical comparison).

Tohoku). Many errors involve orthographic variants such as 障がい (shōgai) 'disability', where partial replacement of kanji with hiragana functions as a mitigation strategy. Instances of 障がい show higher error rates than other items in this domain. Excluding this variant reduces the overall domain error, suggesting that orthographic mitigation poses a distinct challenge for the models.

The *Appearance* and *Workplace & Economy* domains show moderate-to-high error rates, especially in Split 2 (often around 30–40% or higher). In the *Appearance* domain, XLM-R exhibits higher error rates than the monolingual Japanese models, suggesting that culturally grounded or metaphorical expressions may benefit from language-specific pretraining. In both domains, errors are frequently associated with culturally grounded or context-dependent expressions. For example, 社会の窓 (*shakai no mado*; literal meaning: window of society) 'open fly' relies on metaphorical mapping, while 出向 (*shukkō*; literal meaning: go outward) 'forced transfer' can be either literal or euphemistic depending on context. Such expressions show elevated error rates relative to other items and contribute disproportionately to domain-level errors, indicating that metaphor and context dependence remain difficult for the models. Notably, in the *Appearance* domain, the multilingual XLM-R model shows higher error rates than the monolingual models, suggesting that culturally grounded expressions may benefit from language-specific representations.

Overall, the results suggest that models perform well on conventionalized euphemisms but struggle with politically conventionalized, orthographically mitigated, and metaphorical expressions, particularly when lexical cues are absent.

7.2.2. Per Register

We analyze performance across register categories (Formal, Neutral, Informal); see Section 4.3 for details on register annotation. Table 6 reports error rates for baseline and fine-tuned models across Split 1 (seen) and Split 2 (unseen).

A consistent pattern emerges across models: the *Formal* register is the most challenging. Baseline error rates are high (e.g., 67% for Rinna, 51% for Tohoku, and 49% for XLM-R), and although fine-tuning reduces error substantially (e.g., 17% for Rinna in Split 1), performance remains relatively worse than in other registers, especially under unseen conditions (e.g., 29% for Tohoku in Split 2). This suggests that models struggle to capture euphemistic meaning in formal contexts, where indirectness and honorific forms encode pragmatic meaning beyond surface cues.

In contrast, the *Informal* register shows the lowest error rates across models. Fine-tuned models achieve relatively low error (e.g., 12% for XLM-R in Split 1), and baseline performance is also lower than in other registers. However, error increases in Split 2, indicating limited generalization to unseen expressions.

The *Neutral* register shows intermediate but more variable performance. While error is moderate in Split 1, it increases in Split 2 (e.g., up to 27% for Tohoku), suggesting that neutral expressions rely on both lexical and contextual cues and lack consistent surface patterns.

Model differences further highlight sensitivity to register. Tohoku shows higher error in *Formal*, particularly in Split 2, while Rinna remains relatively stable but slightly weaker in *Neutral*. XLM-R exhibits more balanced but less specialized performance across registers.

Overall, these results indicate that euphemism detection depends not only on lexical cues but also on register-specific pragmatic information. Formal expressions remain difficult even after fine-tuning, whereas informal expressions are more easily captured due to their lexical transparency.

7.2.3. Additional Experiment: Tokenization and Pretraining Effects

The higher error rates of Tohoku in the Formal register suggest that performance differences may be partly driven by tokenization and pretraining. While Rinna and XLM-R use SentencePiece, Tohoku relies on MeCab/UniDic-based morphological analysis followed by WordPiece segmentation.

To examine this, we analyzed tokenization fragmentation relative to morpheme boundaries using Sudachi-based segmentation. Formal Japanese expressions often involve *keigo*, which consists of honorific prefixes and auxiliary constructions and thus exhibits complex morphology.

On matched Formal spans shared by all models, Tohoku produced more subword tokens per morpheme than Rinna and XLM-R (paired Wilcoxon tests, $p < 10^{-12}$), with mean differences of +4.5 and +2.4 tokens, indicating greater fragmentation

of morphologically complex expressions. This pattern was particularly evident for keigo forms.

We also observed that Tohoku-specific errors in the matched keigo subset were concentrated in a small number of honorific and euphemistic lexemes (e.g., ご逝去 'passing away', ご年配 'advanced age'). Tohoku further showed higher error rates on keigo expressions compared to the other models.

These findings suggest that Tohoku's elevated error in the Formal register is partly related to tokenization. More fine-grained segmentation may fragment semantically coherent honorific units, making them harder to interpret. However, given the limited data, these results should be interpreted as suggestive rather than conclusive.

8. Discussion

The results provide several insights into how current models interpret Japanese euphemisms.

Model performance is strongly influenced by lexical familiarity. When the same expressions appear in both the training and test sets, all models perform well; however, performance drops substantially for unseen euphemistic forms. This pattern suggests that models rely on memorized lexical patterns rather than inferring the underlying pragmatic functions of euphemisms. The limitation is particularly evident in context-dependent cases, where the distinction between literal and euphemistic meaning depends on cues beyond the sentence. In domains such as *Workplace and Economy*, models frequently misclassify literal descriptions as euphemistic or fail to detect softened references, indicating difficulty in leveraging broader contextual information.

Models also struggle with euphemisms that overlap with formality, politeness, or metaphor. Political euphemisms, for instance, are highly conventionalized and tied to specific genres, leading models to treat them as fixed polite formulas rather than softened criticism, as in 遺憾 (ikan) 'regrettable'. A similar pattern appears with metaphor-based expressions. In the *Appearance* domain, expressions such as 社会の窓 (shakai no mado) 'open fly' rely on visual metaphor to mitigate embarrassment, which requires interpretation beyond the literal meaning. These results suggest that metaphor and pragmatic softening remain challenging for current models.

Orthographic variation introduces an additional challenge. Expressions such as 障がい (shōgai) 'disability', which avoid standard kanji, are often used to convey a gentler impression. However, the models do not appear to treat such variation as a meaningful cue.

Further analysis suggests that tokenization may contribute to model differences, particularly for Tohoku BERT. Honorific expressions (keigo) in Japanese are morphologically complex and encode pragmatic meaning through prefixes and auxiliary constructions. Using Sudachi-based segmentation (Takaoka et al., 2018), we observed that Tohoku produces higher token-to-morpheme ratios on matched formal-register spans, indicating greater fragmentation of these expressions. Errors unique to Tohoku are also concentrated in keigo-related lexemes. While this suggests that segmentation may affect the representation of pragmatically marked forms, the current dataset is insufficient to establish a direct causal relationship.

Finally, XLM-R achieves the strongest performance, particularly in the unseen setting, suggesting that multilingual pretraining may provide broader semantic coverage or exposure to paraphrastic variation. However, in domains involving culturally grounded expressions, monolingual models sometimes perform comparably or better, indicating that language-specific representations remain important for capturing pragmatic nuance.

9. Conclusions

This study examined how transformer-based models interpret Japanese euphemisms under controlled data-splitting conditions designed to isolate lexical familiarity, generalization, and pragmatic inference. The analysis provides a more precise picture of where current systems succeed and where they fail. The results demonstrate that model performance is systematically shaped by domain, register, and pragmatic complexity.

In particular, supplementary analyses suggest that differences in tokenization and morphological segmentation may contribute to variation in performance on formal expressions, especially those involving honorific and euphemistic forms. While these effects remain tentative, they highlight the importance of considering linguistic representation strategies alongside model architecture and training data.

Future work should investigate why multilingual models generalize better than Japanese-only models. This may involve comparing pretraining corpora, analyzing tokenization behavior, or evaluating how deeply each model captures contextual meaning. Examining attention patterns may also help clarify how models process orthographic variation and formulaic expressions.

Another important direction is to extend the dataset to euphemisms that overlap with idioms (e.g., potentially idiomatic expressions (PIEs)), politeness formulas, and honorific forms, allowing systematic comparison between pragmatic and figurative non-literal meaning. Many errors appear

to arise when euphemisms interact with formal or conventionalized linguistic patterns, so incorporating these pragmatic categories would allow a more fine-grained analysis of how different cues influence model behavior.

Finally, adding longer context or document-level data would allow models to distinguish literal and euphemistic readings in ambiguous cases. This is especially important for domains such as Workplace and Economy, where sentence-level context is often insufficient to determine whether an expression softens an unwelcome event.

10. Limitations

A key limitation of this study is the relatively small dataset size and uneven distribution across domains and PET types. While this reflects the natural distribution of euphemisms in corpora, it restricts the ability to draw strong conclusions about domain- or register-specific effects. Future work with larger and more balanced datasets will be necessary to validate these patterns.

Acknowledgements

We would like to thank the Montclair State University NLP Lab for providing the opportunity to conduct this research. This work builds on the lab' s prior studies on euphemism analysis. We also thank the NLP Lab members for their helpful feedback and discussions.

Thank you Julia Sammartino and Patrick Lee for laying the baselines for much of our lab's euphemism work!

11. Bibliographical References

Keith Allan and Kate Burridge. 1991. *Euphemism and dysphemism: Language used as shield and weapon*. Oxford University Press.

Keith Allan and Kate Burridge. 2006. *Forbidden words: Taboo and the censoring of language*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*, 2 edition. Cambridge University Press, Cambridge.

Hasan C. Biyik, Patrick Lee, and Anna Feldman. 2024. Turkish delights: A dataset on turkish euphemisms. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 71–80, Bangkok, Thailand and Online. Association for Computational Linguistics.

Miguel Casas Gómez. 2009. Towards a new approach to the linguistic definition of euphemism. *Language Sciences*, 31:725–739.

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 ACL*.

Haruko Minegishi Cook. 2006. Japanese politeness as an interactional achievement: Academic consultation sessions in japanese universities. *Multilingua*, 25(3):269–291.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Shoko Hamano. 1998. *The Sound-Symbolic System of Japanese*. CSLI Publications, Stanford, CA.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Patrick Lee, A. Chirino Trujillo, D. Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. Meds for pets: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881. Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, pages 184–190. Association for Computational Linguistics.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic

terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of NAACL-HLT 2022*, pages 4437–4452. Association for Computational Linguistics.

Senko K. Maynard. 1997. *Japanese communication: Language and thought in context*. University of Hawai'i Press.

Steven Pinker. 2003. *The blank slate: The modern denial of human nature*. Penguin Books.

Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Viking, New York.

Hussein Rababah. 2014. The translatability and use of x-phemism expressions in medical discourse. *Studies in Literature and Language*, 9:1–12.

rinna Co., Ltd. 2023. rinna japanese roberta model. https://huggingface.co/rinna/japanese-roberta-base.

Kazuma Takaoka, Hiroki Ouchi, Toshinori Sakai, Satoshi Akiyama, and Takashi Yamada. 2018. Sudachi: A japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tohoku University. 2020. cl-tohoku/bert-base-japanese-v2. https://github.com/cl-tohoku/bert-japanese.

James Valentine. 1998. Naming the other: Power, politeness and the inflation of euphemisms. *Sociological Research Online*, 3(4):37–53.

A. Appendix

A.1. Per-PET statistics

Statistic	Mean (M)	SD	Min	Max
Sentences per PET	16.7	11.3	3	68
Euphemistic instances	10.8	4.8	1	39
Non-euphemistic instances	5.9	8.8	0	45
PET ambiguity entropy (bits)	0.45	0.48	0.0	1.0
Tokens per sentence†	58.7	13.5	–	–
Lexical density†	0.79	0.06	–	–

Note: Ambiguity was computed as $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$.
† Approximate corpus-level averages.

A.2. Test-set distribution

Domain	PET Type	Label 0	Label 1
Aging	always	0	40
	sometimes	4	4
Appearance	always	0	40
	sometimes	7	4
Bodily Functions	always	0	87
	sometimes	23	24
Crime & Social Issues	always	0	14
	sometimes	18	15
Death & Dying	always	0	158
	sometimes	28	22
Illness & Disability	always	—	—
	sometimes	14	15
Others	always	0	21
	sometimes	3	4
Personality	always	0	36
	sometimes	5	7
Politics & War	always	0	30
	sometimes	—	—
Sexuality & Relationships	always	0	81
	sometimes	17	15
Workplace & Economy	always	0	31
	sometimes	7	5

Note: Counts for sometimes-euphemistic PETs are approximately label-balanced; minor deviations reflect PETs with uneven label distributions.

A.3. Per-label F1 by PET type

Model	Split 1 (lexical overlap)			Split 2 (no overlap)		
	Sometimes		Always	Sometimes		Always
	$F1_1$	$F1_0$	$F1_1$	$F1_1$	$F1_0$	$F1_1$
Rinna RoBERTa	0.75	0.62	0.91	0.64	0.40	0.93
Tohoku BERT	0.82	0.78	0.90	0.63	0.50	0.88
XLM-RoBERTa	0.84	0.80	0.89	0.68	0.56	0.88

Note. Always-euphemistic PETs contain only label 1 instances by definition; therefore, $F1_0$ is not applicable for the always-only subset.

Steering Pragmatic Interpretation in LLMs: A Diagnostic Evaluation of Few-Shot and Reasoning-Based Prompting for Indirect Speech Acts

Massimiliano Orsini[1], Dominique Brunato[2]
[1]University of Padua, Padua, Italy
[2] Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa, Italy
massimiliano.orsini10@gmail.com, dominique.brunato@ilc.cnr.it

Abstract

Pragmatic competence presents a persistent challenge for Large Language Models (LLMs), as it requires context-dependent inference beyond literal meaning. This study examines whether few-shot prompting can reliably steer LLMs toward appropriate interpretations of indirect speech acts under small-data conditions. Focusing on Italian, we evaluate three LLMs on a small dataset that captures pragmatic ambiguity through graded plausibility judgments. We compare a zero-shot baseline with multiple few-shot prompting configurations that vary in the number and composition of demonstrations, as well as in the presence of explicit pragmatic guidance. Results show that few-shot prompting does not yield robust or monotonic improvements overall. While performance improves substantially for conventionalized indirect speech acts, gains for non-conventionalized indirect speech acts are unstable and limited. In contrast, introducing explicit pragmatic reasoning along with demonstrations through guided chain-of-thought prompting appears more promising. Overall, these findings highlight the limits of example-based steering for pragmatic inference and suggest that explicitly modeling pragmatic reasoning may be a more effective direction in small-data settings.

Keywords: Indirect Speech Acts, Few-shot Prompting, Large Language Models Evaluation

1. Introduction

Pragmatic competence is a fundamental component of human communication. Beyond the literal meaning of sentences, speakers routinely convey intentions that depend on contextual inference, shared knowledge, and socially grounded expectations. Modeling such phenomena remains a major challenge for natural language processing systems, as it requires sensitivity not only to lexical and syntactic information, but also to discourse context and inferential reasoning.

Recent advances in large language models (LLMs) have led to substantial improvements across a wide range of linguistic tasks. Nevertheless, pragmatic phenomena remain especially hard, precisely because they rely on context-dependent and socially embedded knowledge rather than stable form–meaning mappings (Ma et al., 2025). In addition to modeling challenges, the study of pragmatic competence faces a methodological bottleneck. Constructing reliable resources for training and evaluation is non-trivial, and datasets targeting pragmatics are often limited in size or focused on specific, easily operationalizable phenomena. Moreover, evaluation is complicated by human label variation, which is particularly pronounced in pragmatic tasks (Jiang and Marneffe, 2022). These factors make it difficult to assess whether model behavior reflects genuine pragmatic understanding or superficial pattern matching.

Among the various pragmatic phenomena, indirect speech acts (ISAs) constitute a particularly well-studied and theoretically grounded case. Indirect speech acts are utterances whose intended communicative function diverges from their literal form (Searle, 1979). Correctly interpreting them requires recovering the speaker's intention by integrating linguistic form with contextual cues, rather than relying on surface meaning alone. A further crucial distinction concerns the difference between conventionalized and non-conventionalized ISAs. In conventionalized ISAs (C-ISAs), specific lexical or syntactic forms are strongly associated with particular communicative intentions (e.g. "*Can you...?*" for indirect requests) to the point that they induce a strong bias toward the indirect meaning in speakers, making it difficult to access to the literal interpretation even when the context requires it (Gibbs, 1983; Marocchini and Domaneschi, 2022). In contrast, non-conventionalized ISAs (NC-ISAs) derive their indirectness from context-specific reasoning rather than established linguistic conventions (Trott and Bergen, 2019; Bašnáková et al., 2013). Accordingly, while C-ISAs may benefit from distributional regularities learned during pretraining, NC-ISAs require more flexible contextual generalization and inference.

In our previous work, we provided empirical evidence supporting the theoretical distinction between conventionalized and non-conventionalized ISAs (Orsini and Brunato, 2025). That study focused on the construction of a small, manually curated benchmark for Italian and on the analysis of human plausibility judgments, with the goal of capturing pragmatic ambiguity across ISA

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 12–20
11 May 2026.

types. We also reported preliminary results on the performance of Italian LLMs, observing a clear advantage on conventionalized ISAs over non-conventionalized ones. However, model evaluation was not the primary focus of that work, leaving open questions regarding the robustness of model behavior under different evaluation settings and prompting conditions. Building on this previous work, we move beyond this initial analysis and focus on a more systematic evaluation of LLMs' pragmatic competence, specifically investigating the role of few-shot prompting as a lightweight steering mechanism.

Rather than aiming to improve performance per se, our goal is diagnostic: we ask whether few-shot prompting provides a reliable and stable way to steer pragmatic interpretation in a small-data regime, and whether any gains generalize beyond pattern-based cases.

Specifically, we address the following research questions:

- RQ1. To what extent can prompting-based interventions steer LLMs' pragmatic interpretation of ISAs beyond a zero-shot baseline?
- RQ2. Do prompting effects differ across conventionalized indirect speech acts, non-conventionalized indirect speech acts, and literal scenarios?
- RQ3. Does introducing explicit pragmatic knowledge and reasoning strategies in the prompt (e.g. via chain-of-thought instructions) lead to more stable and interpretable improvements in indirect speech act interpretation, compared to example-based prompting alone?

To answer these questions, we evaluate three Italian-capable LLMs under a range of zero-shot and few-shot prompting conditions, using both strict and relaxed accuracy measures derived from graded plausibility scores. Our results show that few-shot prompting yields non-monotonic and fragile effects, improving performance on conventionalized cases while failing to stabilize non-conventionalized interpretation and consistently reducing accuracy on literal scenarios.

2. Related Works

The automatic identification of Indirect Speech Acts (ISAs) has been explored within broader efforts to benchmark pragmatic competence in large language models. While evaluation resources for syntax and semantics are now abundant, pragmatic phenomena — and ISAs in particular — remain underrepresented in standard benchmark suites. This imbalance reflects both conceptual and methodological challenges: unlike morpho-syntactic or lexical-semantic tasks, ISA recognition requires modeling the mismatch between sentence form and communicative function, a phenomenon that is highly context-dependent and often inherently ambiguous.

Among existing resources, many attempts to scale ISA evaluation focus on relatively constrained and structurally regular subclasses of indirectness. Large-scale benchmarks such as BIG-Bench (Srivastava et al., 2023), as well as more targeted datasets including CIRCA (Louis et al., 2020) and GRICE (Zheng et al., 2021), which are part of the Pragmatic Understanding Benchmark (Sravanthi et al., 2024), typically tackle indirect speech acts through specific, recurring patterns, most notably indirect answers to polar questions or other highly conventionalized forms of indirect response. These settings allow for comparatively straightforward data generation and annotation, often supporting binary classification (direct vs. indirect) or limited interpretation spaces. While such resources have proven valuable for probing pragmatic inference at scale, their focus on restricted ISA types inevitably narrows the range of indirectness phenomena being modeled, underrepresenting not only more context-sensitive NC-ISAs but also less frequent lexical triggers of C-ISAs.

A different line of work seeks to enrich contextual information rather than expanding dataset size through structural constraints. Hu et al. (2023) propose scenario-based tasks for several pragmatic phenomena, in which models are presented with short contextual descriptions and must select the appropriate interpretation of an utterance among literal, indirect, and distractor options. This design increases contextual variability and more closely approximates natural discourse conditions, explicitly foregrounding the role of pragmatic inference. A closely related approach is adopted by Park et al. (2024), who evaluate LLMs' pragmatic abilities on a small Korean dataset targeting the four Gricean maxims, using a multiple-choice task in which models are given a context and an utterance to interpret among four possible interpretations. More recently, we introduce a resource explicitly focused on ISAs, named INDIR-IT Orsini and Brunato (2025), inspired by the scenario-based methodology of Hu et al. (2023) but narrowing its scope to ISA phenomena alone. Crucially, the dataset distinguishes between non-conventionalized ISAs (NC-ISA), which rely heavily on contextual inference, and conventionalized ISAs (C-ISA), where indirect meanings are more lexically or pragmatically routinized. This distinction allows for a finer-grained investigation of pragmatic competence, separating cases where indirectness emerges from general inferential reasoning from those supported by established communicative conventions. At the same time, the dataset

inherits the core trade-off of expert-designed resources: improved theoretical control and interpretability at the cost of limited scale.

3. Experimental Setting

3.1. The Task

The task follows the original design of the INDIR-IT [1]. Each item consists of a short everyday-life scenario involving two characters. The model is asked to evaluate the communicative intention of one of the characters. For each scenario, the model is presented with four candidate interpretations of the speaker's intended meaning: one indirect meaning, one literal meaning, two lexical distractors.

The dataset consists of three types of scenarios:

- C-scenarios: containing C-ISAs
- L-scenarios: containing C-ISAs whose literal meaning is favored
- NC-scenarios: containing NC-ISAs

Exmples of these three types of scenarios can be found in Appendix A.

Rather than selecting a single label, models are instructed to assign a plausibility score from 1 to 5 to each interpretation, where higher values indicate greater plausibility. This formulation explicitly allows for graded judgments and inherent pragmatic ambiguity, rather than enforcing categorical decisions.

3.2. Prompting Conditions

For the diagnostic purpose of our study, we compare a *zero-shot* baseline against several *few-shot prompting conditions* designed to test whether limited exposure to ISAs can steer model behavior, and if this effect is consistent across different types of ISAs. The conditions include:

- zero-shot (0-shot): task instructions only;
- C-scenarios-only, 5 and 10 shots (C5, C10).
- NC-scenarios-only, 5 and 10 shots (NC5, NC10).
- L-scenarios-only, 5 shots (L5)
- Pairs of C-scenarios and L-scenarios sharing the same utterance, 5 shots (CL).
- Mixed C-scenarios and NC-scenarios, 5 and 10 shots (M5, M10).
- Mixed enhanced (ME): This condition mirrors the M10 prompt, but includes an additional instruction explicitly warning the model that the speaker's communicative intention may diverge from the literal meaning of the utterance.
- Guided chain-of-thoughts (CoT): where models were provided with a four-step heuristic inference process and 5 demonstrations where this process is applied. The steps require the model to first identify the literal meaning of the utterance among the four options, then to detect contextual cues that may confirm or reject this interpretation, and finally to select the appropriate option once a decision has been reached. The four steps are illustrated in Appendix B.

In all few-shot conditions, demonstration examples are selected following the design principles of the original dataset. In particular, examples are chosen so as to preserve, as far as possible, the distribution of combinations between primary acts (intended meanings, e.g. request) and secondary acts (surface forms, e.g. question) observed in the resource. This choice is meant to avoid introducing systematic biases in the demonstrations and to ensure that few-shot prompting reflects the diversity of pragmatic mappings present in the data.

The ratings for each interpretation are derived from the average of multiple human annotations collected as part of the INDIR-IT release. Since the task requires discrete scores on a 1–5 scale, these averaged values are rounded to the nearest integer before being included in the prompt.

In Table 1, we show the basic instructions common to all the prompting conditions.

3.3. Models

We evaluate three large language models that differ in training data and optimization strategies:

- LLaMA-3.1 (8B-Instruct)[2]: this model is the most compact variant in the Llama 3.1 series, featuring an instruction-tuned architecture optimized for multilingual tasks and high-efficiency inference (Dubey et al., 2024).
- ANITA[3]: an Italian-specific adaptation of Llama 3, this model was developed using QLoRA for parameter-efficient fine-tuning and Direct Preference Optimization (DPO) to align its outputs with regional linguistic standards (Polignano et al., 2026).

[1] The resource is available on HuggingFace at the following link: https://huggingface.co/datasets/MaxiOr/INDIRIT

[2] HuggingFace handle: meta-llama/Llama-3.1-8B-Instruct

[3] HuggingFace handle: swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

Prompt instructions	English translation of the instructions
COMPITO: Leggerai delle storie brevi che descrivono una situazione ordinaria tra due personaggi: Fausto e Margherita. Ogni storia si conclude con una frase che Fausto rivolge a Margherita. Per ogni storia vengono fornite quattro possibili interpretazioni per spiegare l'intenzione comunicativa della frase di Fausto, in relazione alla situazione presentata. Ad ogni interpretazione, dovrai assegnare un punteggio da 1 a 5 in base alla sua plausibilità: (1 = non plausibile, 2 = poco plausibile, 3 = plausibile, 4 = più che plausibile, 5 = molto plausibile). Fornisci esclusivamente i punteggi finali per ciascuna interpretazione, senza spiegazioni o passaggi intermedi. --- SHOTS --- Non scrivere spiegazioni. Non scrivere testo aggiuntivo. Qualsiasi cosa fuori dal formato è un errore.	TASK: You will read short stories describing an ordinary situation between two characters: Fausto and Margherita. Each story ends with a sentence uttered by Fausto and addressed to Margherita. For each story, four possible interpretations are provided to explain the communicative intention of Fausto's sentence, in relation to the situation presented. For each interpretation, you must assign a score from 1 to 5 based on its plausibility: (1 = not plausible, 2 = slightly plausible, 3 = plausible, 4 = more than plausible, 5 = very plausible). Provide only the final scores for each interpretation, without explanations or intermediate steps. --- SHOTS --- Do not write explanations. Do not write additional text. Anything outside the required format is an error.

Table 1: Excerpt of the prompt instructions used in the experimental setup, along with their English translation.

- Minerva (7B)[4]: the instruct-tuned version of the largest entry in its specific family of LLMs and is natively trained in the Italian language (Orlando et al., 2024).

All models are accessed through the Hugging Face inference API and queried in a chat-based setting. We do not perform any parameter updates; all experiments are conducted in a prompting-only regime. No sampling strategy was employed at inference time and decoding was fully deterministic.

3.4. Evaluation Metrics

We calculated two accuracy measures[5]: **Strict accuracy**, where a prediction is counted as correct only if the target interpretation receives a strictly higher score than all other options. **Relaxed accuracy**, where a prediction is counted as correct if the target interpretation is tied for the highest score.

Relaxed accuracy is particularly appropriate for pragmatic tasks, where multiple interpretations may remain plausible even for human annotators. This metric allows us to distinguish between outright misinterpretations and cases where models recognize the intended meaning but fail to confidently separate it from alternatives.

[4]HuggingFace handle: sapienzanlp/Minerva-7B-instruct-v1.0

[5]Dataset coverage ranges from 90 items in the 10-shot conditions to 100 items in the 0-shot condition.

4. Results

Before analyzing the results, we first examined how the models adhered to the required output format. Llama3.1 consistently followed the format, whereas Anita generally did so, with only a few deviations occurring in prompts with 10-shot demonstrations. Minerva, by contrast, often produced only the interpretation it deemed correct, rather than providing all interpretations with their scores. In cases where the format was not respected, responses were still counted and evaluated if the intended answer could be inferred; otherwise, they were marked as incorrect.

As shown in Table 2, across all models, **few-shot prompting does not yield consistent improvements** over the zero-shot baseline. Both strict and relaxed accuracy exhibit **non-monotonic behavior** with respect to the number of demonstrations: in several conditions, performance improves with 5-shot prompting but stagnates or degrades when increasing to 10 shots. A notable exception is the NC10 condition, which yields comparatively better results for both ANITA and Minerva, as better shown in Figure 1.

These aggregate trends are largely reflected at the level of individual models, suggesting that few-shot effects are fragile and highly sensitive to configuration. In particular, this degradation may be partially attributed to increased prompt length, which has been shown to negatively affect model performance beyond a certain threshold (Levy et al., 2024)

Considering performance by scenario type, **in**

C	ALL	C	L	NC
0S	0.57 / 0.26	0.74 / 0.38	0.54 / 0.21	0.46 / 0.20
C5	0.51 / 0.31	0.73 / 0.56	0.39 / 0.22	0.46 / 0.23
C10	0.48 / 0.30	0.75 / 0.57	0.36 / 0.19	0.39 / 0.23
NC5	0.48 / 0.33	0.67 / 0.51	0.39 / 0.23	0.41 / 0.25
NC10	0.47 / 0.39	0.62 / 0.59	0.42 / 0.34	0.37 / 0.24
L5	0.47 / 0.30	0.61 / 0.44	0.37 / 0.23	0.43 / 0.23
CL	0.41 / 0.32	0.61 / 0.53	0.21 / 0.16	0.41 / 0.29
M5	0.48 / 0.32	0.71 / 0.52	0.36 / 0.21	0.41 / 0.25
M10	0.46 / 0.33	0.57 / 0.57	0.37 / 0.26	0.38 / 0.23
ME	0.46 / 0.34	0.67 / 0.56	0.36 / 0.27	0.40 / 0.26
CoT	0.55 / 0.39	0.69 / 0.55	0.62 / 0.43	0.40 / 0.24

Table 2: Relaxed and strict accuracy (R / S) averaged across three LLMs (LLaMA 3.1, ANITA, Minerva), broken down by prompting condition (rows) and scenario type (columns).

the C-scenarios, all models show a substantial improvement over the zero-shot baseline under most conditions. Accuracy rises sharply and approaches ceiling values, with minimal differences between strict and relaxed evaluation.

This indicates that conventionalized cases are reliably learnable from surface-level patterns and benefit from limited exposure, independently of the specific few-shot configuration.

As for the **L-scenarios, no clear or consistent trend emerges across models**. The only notable effect is observed for Llama3.1, which shows a marked improvement in both relaxed and strict accuracy for literal interpretations under the CoT prompt, without any degradation in C or NC scenarios. For the other models, performance remains highly variable across prompting conditions and does not exhibit systematic improvements.

Lastly, in the **NC-scenarios**, for relaxed accuracy, **some few-shot configurations yield small local improvements** over zero-shot performance, but these gains are inconsistent and not cumulative. Increasing the number of demonstrations does not stabilize performance, and different configurations favor different models. On the other hand, for strict accuracy, all prompts improve, although the zero-shot baseline was already very low for the models (LLama3.1: 0.3, Anita: 0.1, Minerva: 0.2).

As further illustrated in Figure 1, relaxed and strict accuracy generally follow the same qualitative trends observed in the aggregated results. LLaMA 3.1 and Minerva exhibit comparable performance across both metrics, whereas ANITA shows a more pronounced gap between relaxed and strict accuracy. This reflects a higher frequency of tied scores among candidate interpretations, suggesting that ANITA more often distributes plausibility across multiple options rather than sharply favoring a single interpretation.

Overall, no single prompting factor emerges as a reliable determinant of performance across conditions. Instead, results suggest that task difficulty plays a central role: simpler items—such as those involving more frequent lexical triggers (e.g. *Puoi V?/Can you V?*) or primary/secondary act pairings (e.g stative act as positive/negative response) consistently achieve higher accuracy rates, regardless of the prompting strategy.

Regarding the CoT prompt, one possible explanation is that it facilitates the identification of literal interpretations, which may account for the observed improvements in the L-scenario condition.

It is also plausible that the CoT formulation shifts the nature of the task, making it closer to a classification problem rather than to a NLI task. This could explain why improvements remain limited in the NC condition, where indirect and literal interpretations are not in competition but structurally dependent, as the indirect interpretation is only licensed if the literal one is contextually valid.

5. Discussion and Conclusion

This study examined the impact of few-shot prompting on the interpretation of indirect speech acts in Italian, considering conventionalized, non-conventionalized, and literal scenarios.

Concerning the extent to which prompting-based interventions can steer the models' pragmatic interpretation beyond a zero-shot baseline, the results show that few-shot prompting generally does not provide robust or systematic improvements, with performance varying substantially across configurations and often degrading as more demonstrations are added.

As for the differences across conventionalized and non-conventionalized ISAs, as well as literal meaning of conventionalized ISAs, prompting effects clearly vary: only C-ISAs appear to benefit from in context learning from all the tested conditions, and guided chain-of-thought prompting also succeeds in mitigating the bias toward indirect meanings and improves performance on literal scenarios, while also preserving strong results on conventionalized cases. However, none of the strategies produces measurable gains on non-conventionalized indirect speech acts, which remain inherently difficult to construct and reason about.

Lastly, through the ME and CoT prompting conditions, we examined whether introducing explicit pragmatic knowledge and reasoning in the prompts could enhance model performance. Our results showed that the CoT prompt emerges as promising only for Llama 3.1, the most robust model in our evaluation: by explicitly structuring the inferential process, it improves performance without negatively affecting other scenarios. Importantly, this improvement in the L-scenarios should not be underestimated, as the literal interpretation of C-ISAs has been shown to be highly challenging even for human subjects (Gibbs, 1983). This suggests

Figure 1: Models' performance (relaxed and strict accuracy) across all conditions and scenarios.

that, at least for stronger models, explicit reasoning guidance-rather than additional examples—may be key to enhancing pragmatic interpretation in large language models.

However, despite avoiding performance degradation on NC-scenarios, none of the prompting strategies leads to measurable improvements in this setting. This is particularly problematic, as non-conventionalized indirect speech acts are inherently harder to construct and therefore difficult to scale into large datasets. In this context, the ability to learn from small numbers of examples would be especially desirable. One possible direction for future work is to further enrich the guided inferential process, by expanding the number and specificity of reasoning steps and explicitly directing models to attend to fine-grained contextual cues, while simultaneously reducing the number of examples in order to mitigate degradation due to prompt length, and possibly, instructing the model to output its reasoning in order to enhance interpretability for further analysis.

6. Acknowledgements

This work has been (partially) supported by XAI-CARE - PNRR-MAD-2022-12376692 project, under the NRRP MUR program funded by the NextGenerationEU and by LLMs4EU "Large Language Models for the European Union" project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) under the grant agreement 101198470.

7. Bibliographical References

Jana Bašnáková, Kirsten Weber, Karl Magnus Petersson, Jos van Berkum, and Peter Hagoort. 2013. Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex*, 24(10):2572–2578.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Raymond W Gibbs. 1983. Do people always process the literal meanings of indirect requests? *Journal of experimental psychology. Learning, memory, and cognition*, 9(3):524–533.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Eleonora Marocchini and Filippo Domaneschi. 2022. "can you read my mind?" conventionalized indirect requests and theory of mind abilities. *Journal of Pragmatics*, 193:201–221.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Massimiliano Orsini and Dominique Brunato. 2025. Direct and indirect interpretations of speech acts: evidence from human judgments and large language models. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 837–848.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. Pragmatic competence evaluation of large language models for the korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2026. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Scientific Reports*, 16.

John R. Searle. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, "...", and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Sean Trott and Benjamin Bergen. 2019. Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, 56(8):675–707.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

8. Appendices

A. Examples of Non-conventional Scenarios and Liter/Conventional Pairs

NC-scenario:

Margherita chiede a Fausto se sa se una loro conoscente sia sposata. Fausto le dice: "Le ho visto un anello al dito."

I - Fausto conferma a Margherita che la loro conoscente sia sposata.
L - Fausto informa Margherita che la loro conoscente possiede un anello.
D1 - Fausto vuole dire a Margherita che non sa se la loro conoscente sia sposata ma che di sicuro è molto ricca.

D2 - Fausto vuole dire che a lui non importa se la loro conoscente sia sposata.

English translation:

Margherita asks Fausto if he knows if an acquaintance of theirs is married. Fausto tells her: "I saw a ring on her finger."

I - Fausto confirms to Margherita that their acquaintance is married.
L - Fausto informs Margherita that their acquaintance owns a ring.
D1 - Fausto wants to tell Margherita that he doesn't know if their acquaintance is married but that she is certainly very wealthy.
D2 - Fausto wants to tell her that he doesn't care if their acquaintance is married.

C/L-pair:

C-scenario:

Margherita e Fausto stanno andando in macchina al supermercato. Fausto a un certo punto chiede a Margherita: "Puoi svoltare a destra?"

L-scenario:

Margherita e Fausto stanno andando in macchina al supermercato. Fausto, che sa che in questo periodo ci sono molte cantieri aperti in strada, a un certo punto chiede a Margherita: "Puoi svoltare a destra?"

Options:

I - Fausto vuole che Margherita svolti a destra.
L - Fausto vuole sapere se si può svoltare a destra.
D1 - Fausto vuole sapere se Margherita è in grado di usare lo sterzo.
D2 - Fausto vuole che Margherita lo riaccompagni a casa.

English translation:

C-scenario:

Margherita and Fausto are going to the supermarket by car. At one point, Fausto asks Margherita: "Can you turn right?"

L-scenario:

Margherita and Fausto are going to the supermarket by car. Fausto, who knows there are many open roadworks at the moment, asks Margherita: "Can you turn right?"

Options:

I - Fausto wants Margherita to turn right.
L - Fausto wants to know if it is possible to turn right.
D1 - Fausto wants to know if Margherita is able to use the steering wheel.
D2 - Fausto wants Margherita to take him back home.

B. Inferential Steps for the Guided CoT

Instructions:

1. Individua quale delle quattro interpretazioni corrisponde al significato letterale della frase.
2. Individua l'indizio contestuale rilevante per trovare l'intenzione finale.
3. Valuta se l'interpretazione letterale coincide anche con l'intenzione comunicativa finale di Fausto.
4. Quale interpretazione rappresenta l'intenzione finale di Fausto?

STORIA:

Mentre Margherita legge un libro sul divano, Fausto cerca di appendere un quadro al muro, ma sta avendo un po' di difficoltà. Fausto dice: "Mi piacerebbe che qualcuno mi aiutasse."

Cosa intende dire Fausto?

a) Fausto esprime a Margherita il desiderio che qualcuno venga ad aiutarlo ad appendere il quadro.
b) Fausto vuole che Margherita gli dia una mano ad appendere il quadro.
c) Fausto vuole distrarre Margherita dalla lettura.
d) Fausto avvisa Margherita che non riuscirà ad appendere il quadro a meno che qualcuno non lo aiuti.

PASSAGGI:

1. Interpretazione letterale?
a) Fausto esprime a Margherita il desiderio che qualcuno venga ad aiutarlo.
2. Indizio contestuale?
Margherita ha la possibilità di aiutare Fausto.
3. Interpretazione letterale è l'intenzione finale? No.

4. Intenzione finale? b) Fausto vuole che Margherita gli dia una mano ad appendere il quadro.

English translation:

1. Identify which of the four interpretations corresponds to the literal meaning of the sentence.
2. Identify the relevant contextual cue to determine the final intention.
3. Assess whether the literal interpretation also matches Fausto's final communicative intention.
4. Which interpretation represents Fausto's final intention?

STORY:
While Margherita is reading a book on the couch, Fausto is trying to hang a picture on the wall, but he is having some difficulty. Fausto says: "I would like someone to help me."

What does Fausto mean?
a) Fausto expresses to Margherita the desire that someone comes to help him hang the picture.
b) Fausto wants Margherita to help him hang the picture.
c) Fausto wants to distract Margherita from reading.
d) Fausto informs Margherita that he will not be able to hang the picture unless someone helps him.

STEPS:
1. Literal interpretation? a) Fausto expresses to Margherita the desire that someone helps him.
2. Contextual cue? Margherita is in a position to help Fausto.
3. Does the literal interpretation match the final intention? No.
4. Final intention? b) Fausto wants Margherita to help him hang the picture.

Injecting Structured Lexicographic Knowledge into LLMs for Non-Literal Expression Disambiguation: A Controlled Study on Croatian

Slobodan Beliga[1,2], Ivana Filipović Petrović[3], Ana Meštrović[1,2]
[1]Faculty of Informatics and Digital Technologies, University of Rijeka, Rijeka, Croatia
[2]Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Rijeka, Croatia
[3]Croatian Academy of Sciences and Arts, Zagreb, Croatia
sbeliga@inf.uniri.hr, ifilipovic@hazu.hr, amestrovic@inf.uniri.hr

Abstract

In potentially idiomatic expressions (PIEs), the same surface form may receive either a literal or an idiomatic interpretation depending on context, making automatic literal–idiomatic disambiguation challenging. This is acute for Croatian, where annotated data and locally runnable generative models are limited. We present a study of Croatian PIE literal–idiomatic disambiguation examining how structured lexicographic knowledge can improve open-weight, decoder-only LLMs without fine-tuning. Using a new expert-annotated concordance dataset – CroPIEs, we compare baseline prompting to inference-time knowledge injection via retrieval-augmented generation (RAG) from a Croatian phraseological dictionary. We isolate the contribution of three knowledge types: definitional knowledge (structured meanings), contextual knowledge as curated prototypical usage examples, and their combination. Results show consistent improvements in macro-F1 for both GaMS-2B-Instruct and GaMS-9B-Instruct models. Definitional knowledge is generally more stable than examples alone, while examples can be effective but less consistent across expressions. The strongest and most reliable gains are obtained when definitions and examples are combined, indicating a synergistic effect between explicit meaning descriptions and contextual cues. Per-class analyses show that injected lexicographic evidence mitigates baseline biases between LITERAL and IDIOMATIC predictions, improving decision balance in a low-resource setting with small data of compact, expert-curated lexicographic evidence injected at inference time.

Keywords: literal–idiomatic disambiguation, PIEs, idioms, structured lexicographic knowledge, knowledge injection, LLM, RAG, GaMS

1. Introduction

Non-literal expressions (NLEs), including many phraseological units (PUs), have long been recognized as a persistent challenge for natural language processing(Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017). This tension is especially visible when the same surface string admits both a literal and a figurative reading, making context-sensitive semantic disambiguation unavoidable. A central task in this area is *literal–idiomatic usage disambiguation*: determining whether a given occurrence of a PU is used compositionally or figuratively.

In corpus settings, many candidates can occur either idiomatically or literally depending on context. Following Haagsma et al. (2020), we adopt the term *potentially idiomatic expressions* (PIEs), defined as expressions that *can* have an idiomatic meaning regardless of whether they actually have that meaning in a given context. In other words, the same surface form may receive either a compositional (literal) reading or a non-compositional (idiomatic/figurative) reading depending on contextual cues. Literal–idiomatic disambiguation is therefore a context-sensitive decision that affects meaning preservation in downstream applications such as machine translation, summarization, and information extraction, where misinterpreting idioms can distort meaning.

This decision remains brittle for large language models (LLMs) in practice. Literal occurrences are frequently over-labeled as idiomatic (Tedeschi et al., 2022), and conversational prompting can induce agreement biases unless the output protocol is tightly constrained (De Luca Fornaciari et al., 2024). Conversely, structured prompts with constrained outputs can yield competitive performance on PIE identification without task-specific fine-tuning (Hashiloni et al., 2025), yet benchmark gains may still rely on spurious cues rather than grounded interpretation (Chakrabarty et al., 2022). For less-resourced languages, dictionaries and phraseological repositories provide compact, vetted semantic knowledge, and definitions and curated examples can be exploited as task guidance (Škvorc and Robnik-Šikonja, 2025).

We therefore study inference-time grounding via retrieval-augmented generation (RAG), which injects external knowledge without additional training. Prior evidence suggests that dictionary-augmented prompting can help, but gains depend on coverage

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 21–30
11 May 2026.

and ambiguity handling (Perak et al., 2024). More broadly, RAG is often more cost-effective than parameter updates, yet its impact hinges on retrieval quality and context construction, motivating controlled evaluations that isolate the contribution of curated external knowledge (Abonizio et al., 2025; Bhushan et al., 2025).

Against this background, we present a controlled study of Croatian PIE disambiguation, focusing on idioms as a salient PU subclass. Croatian remains comparatively low-resourced: high-quality *local* LLMs are scarce, and available models are typically not trained primarily on Croatian but only adapted to it, which makes inference-time grounding especially attractive. Our primary goal is to quantify, in a *small data* setting, how much *structured lexicographic knowledge from a Croatian phraseological dictionary* helps literal–idiomatic disambiguation: specifically, the contribution of *(i)* definitional knowledge, *(ii)* curated prototypical usage contexts (examples), and *(iii)* their synergy when both are injected via RAG. Using a new expert-annotated Croatian concordance dataset (10 idioms; 1,000 instances), we compare baseline prompting to three knowledge variants (definitions, examples, and their combination) under identical prompting and decoding constraints, and check robustness on two smaller open-weight decoder-only GaMS models of markedly different sizes (2B vs. 9B). Our contributions are threefold:

1. We release a new expert-annotated Croatian concordance dataset for literal–idiomatic usage disambiguation (1,000 instances; 10 PIEs/idioms).
2. We conduct a controlled evaluation of inference-time lexicographic knowledge injection via RAG on two Croatian-adapted open-weight decoder-only LLMs (GaMS-2B-Instruct and GaMS-9B-Instruct), comparing baseline prompting to three knowledge variants (definitions, usage examples, and their combination).
3. We provide empirical evidence that small, manually curated phraseological resources can improve literal–idiomatic disambiguation and mitigate class-level prediction biases in a low-resource setting, with gains supported by statistical testing.

2. Related Work

Automatic idiom detection and usage disambiguation has traditionally been framed as a binary classification task distinguishing literal from idiomatic usage of identical surface forms. Neural approaches such as MICE (Škvorc et al., 2022) demonstrated that contextual embeddings (e.g. ELMo, BERT) encode signals sufficient for detecting idiomaticity, including cases involving unseen expressions. These findings confirmed the importance of contextualized representations, while also underscoring challenges related to limited annotated data and generalization across idioms.

More recently, large language models have been evaluated on idiomatic and figurative language understanding in prompting-based setups. The DICE benchmark (Mi et al., 2025) investigates LLM-based idiom comprehension and explanation, focusing on generative interpretation across model sizes and prompting strategies. Broader figurative reasoning datasets such as FLUTE (Chakrabarty et al., 2022) further explore non-literal interpretation and contextual inference. Together, these studies indicate that pretrained models capture substantial figurative knowledge, although performance remains sensitive to task formulation and prompting design.

Within the Croatian context, recent research has examined the role of LLMs in lexicographic and phraseological tasks. Studies have explored LLM-assisted conceptual organisation of idioms and semantic grouping within the *Online Dictionary of Croatian Idioms* (Beliga and Filipović Petrović, 2024; Filipović Petrović and Beliga, 2025), as well as AI- and corpus-based strategies for identifying phraseme constructions through hybrid human–LLM workflows (Beliga and Filipović Petrović, 2025). These developments highlight the growing integration of AI tools, corpus technologies, and structured lexicographic resources in Croatian phraseological research.

3. Data

This section describes the CroPIEs-1k concordance dataset and the structured lexicographic resource used for RAG.

3.1. Concordance Dataset

Corpus Source. The dataset was derived from the Croatian web corpus CLASSLA-web.hr 2.0 CLARIN.SI (2024), compiled from the national *.hr* domain (2021–2024) and available via CLARIN.SI through the NoSketch Engine concordancer. The corpus covers heterogeneous web genres (e.g. news, blogs, forums) and provides morphosyntactic annotation and advanced querying, enabling precise extraction of PIE candidates. Importantly, all instances in our dataset were selected on the basis of attested usage in authentic corpus data (i.e., they are not synthetically generated).

Selection of Phraseological Units. Ten verb-based phraseological units were selected (Table 2) to ensure: (i) attestation in contemporary Croatian,

(ii) identical surface form in literal and idiomatic usage, (iii) sufficient frequency in CLASSLA-web.hr 2.0 (each >1,000 occurrences), and (iv) a clear semantic contrast between compositional (literal) and non-compositional (idiomatic) readings. The requirement of identical lexical form ensures that the task involves semantic disambiguation within the same surface string. In all selected cases, the same lexical sequence may receive either a literal, compositionally interpretable reading or an idiomatic, phraseological interpretation depending on context. The expressions exhibit syntactic variability, including inflectional variation (tense, aspect, agreement), word-order alternations, insertion of modifiers, and clitic placement, reflecting the rich morphosyntactic structure of Croatian. Such variation is an inherent property of phraseological usage in context and does not represent deviation from a canonical form. Disambiguation therefore requires contextual semantic interpretation across naturally occurring structural variants. All expressions are specified in their canonical form (Table 1), while their corpus attestations reflect authentic grammatical realizations. All selected expressions are well established in contemporary usage and display substantial corpus frequency, ensuring that the dataset reflects productive language patterns.

Concordance Extraction. For each expression, concordance lines were extracted in KWIC format with ±100 characters of left/right context. Given the high frequency of each expression (>1,000 occurrences), we used the NoSketch Engine random sampling function to select 100 instances per expression (10×100=1,000 total). This function generates a representative subset of concordance lines while preserving corpus distribution across sources and genres. The choice of 100 instances per expression ensures sufficient contextual variability for reliable idiom-level evaluation while maintaining uniform sample size across expressions. The resulting dataset, CroPIEs-1k, will be released publicly upon acceptance (link in the camera-ready version).

Manual Annotation. All instances were manually annotated by an expert linguist in Croatian phraseology as Literal (compositional) or Idiomatic (phraseological). While inter-annotator agreement was not measured due to the single-annotator setup, annotation decisions followed consistent criteria based on contextual semantic interpretation. The relatively clear distinction between literal and idiomatic usage in the selected dataset reduces the likelihood of systematic ambiguity. Overall, the task was generally straightforward for an expert annotator.

Table 1 confirms near-balance across classes, reducing confounding effects due to class imbalance in evaluation.

Table 1: Class distribution across PU_{1-10} ($n = 100$ per subset; total $N = 1000$ in CroPIEs-1k dataset.)

PU	1	2	3	4	5	6	7	8	9	10	$\mu \pm \sigma$
IDM	48	51	45	50	49	52	47	50	48	51	**49.1 ± 2.1**
LIT	52	49	55	50	51	48	53	50	52	49	**50.9 ± 2.1**

3.2. Lexicographic Resource for RAG

The external knowledge for RAG was drawn from the *Online Dictionary of Croatian Idioms*[1] Croatian Academy of Sciences and Arts (2023), an open-access, corpus-based born-digital resource developed at the Croatian Academy of Sciences and Arts since 2019. Such structured lexicographic resources are particularly valuable for knowledge injection into LLMs, as they provide expert-authored, validated semantic evidence that complements web-derived pretraining data (often including user-generated sources such as Wikipedia and blogs) and can help mitigate knowledge gaps and reduce ungrounded generations.

The dictionary combines manual lexicographic analysis with corpus-supported procedures. Entries were compiled in Lexonomy and are based on systematic corpus examination, including frequency analysis and collocational evidence. All definitions and examples were manually selected and edited by lexicographers to ensure representativeness and semantic precision. Version 2 (2023) contains 563 entries covering 1,165 idioms.

For our RAG pipeline, we exported the dictionary from Lexonomy and converted it to JSONL. We created three parallel collections[2] matching our experimental conditions: **(1) Def** (definitions only), **(2) Exs** (examples only), and **(3) Def+Exs** (definitions+examples). Each JSONL record includes a stable sense identifier and idiom-level metadata to enable deterministic matching and consistent retrieval across conditions.

All ten phraseological units in our concordance dataset are covered by the dictionary. For each instance, the pipeline retrieves the corresponding entry and injects its structured content into the prompt, typically one or two definitions (for polysemy) and around two curated examples (occasionally one or three for variant forms). In this study, the dictionary thus provides a small-scale but high-quality expert resource whose structured semantic information is injected to support disambiguation in a low-data setting.

[1] https://lexonomy.elex.is/frazeoloskirjecnikhr

[2] JSONL collections used for retrieval: GITHUB

Expression (Croatian)	Literal Gloss	Idiomatic Meaning (English)
bacati mrvice	to throw crumbs	to offer small concessions deliberately in order to appease someone
okrenuti leđa	to turn one's back	to abandon or withdraw support
isplivati na površinu	to float to the surface	to become visible or publicly known
dati crveni karton	to give a red card	to remove someone from a political or institutional position
graditi mostove	to build bridges	to promote cooperation or reconciliation
biti u sjeni	to be in the shadow	to remain overshadowed or unnoticed
naletjeti na minu	to run into a mine	to encounter an unexpected hidden problem
biti u komi	to be in a coma	to be in a state of lethargy or inactivity
stati na noge	to stand on one's feet	to recover or regain stability
znati koliko je sati	to know what time it is	to know what is going on; to be aware of the situation

Table 2: Phraseological units included in the dataset with literal glosses and idiomatic meanings.

4. Experimental Setup

We evaluate the impact of injecting structured lexicographic knowledge via RAG on Croatian *literal–idiomatic disambiguation*. We compare a baseline condition without retrieval (prompting only) to three RAG variants, using a controlled, paired evaluation design across model sizes.

4.1. Task Formulation and Evaluation

We formulate the task as binary sentence-level *literal–idiomatic disambiguation*. Given a concordance sentence containing a predefined target idiomatic expression (possibly in morphologically or syntactically varied form), the model predicts whether the expression is used LITERAL (compositional) or IDIOMATIC (non-compositional) in that sentence. For each instance x_i, the model outputs $y_i \in \{\text{LITERAL}, \text{IDIOMATIC}\}$; gold labels were assigned manually at the sentence level. Evaluation is strictly categorical (no graded judgments).

Evaluation. We report macro-averaged F1 (macro-F1). Scores are computed (i) globally over all instances and (ii) at the idiom level, where macro-F1 is computed separately for each idiom subset (100 instances). Idiom-level scores are used for paired comparisons across conditions; we report their mean$\pm$SD across the ten idioms. To analyze class-level effects, we also report precision (P), recall (R), and F1 separately for LITERAL and IDIOMATIC (idiom-averaged). For interpretability, we also report Δ macro-F1 relative to the baseline (RAG–baseline) in tables and figures. Statistical significance between baseline and RAG variants is assessed with the Wilcoxon signed-rank test on idiom-level macro-F1 ($n = 10$ idioms), appropriate for the paired and small-sample design. All conditions are evaluated on the same fixed set of instances. A safeguard mechanism was implemented to handle structurally invalid outputs, but no such cases occurred in the reported experiments.

4.2. Models

In this study, we evaluate two open-weight multilingual LLMs from the GaMS (Generative Model for Slovene) family: *GaMS-2B-Instruct*[3] and *GaMS-9B-Instruct*[4]. Both are decoder-only Transformer models based on the Gemma 2 architecture and were continually pretrained and subsequently instruction-tuned for South Slavic languages (Vreš et al., 2024; Vajda et al., 2025). We focus on small, locally runnable open-weight models that can be executed without external APIs on modest GPU hardware, enabling controlled and reproducible experimentation.

The variants differ primarily in scale: GaMS-2B-Instruct has 26 layers (hidden size 2304; 8 attention heads), while GaMS-9B-Instruct has 42 layers (hidden size 3584; 16 attention heads). Both support a maximum context length of 8192 tokens.

GaMS models are not Croatian-specific. While large proprietary LLMs can perform well on Croatian, Croatian remains comparatively low-resourced in terms of openly available, locally runnable instruction-tuned decoder-only LLMs with substantial Croatian coverage. Although primarily developed for Slovene, GaMS was continually pretrained on multilingual corpora that include Croatian (alongside Slovene, Serbian, and Bosnian), making it a practical open-weight choice for controlled experiments on Croatian literal–idiomatic disambiguation.

Comparing 2B and 9B within the same model family keeps the architecture and training paradigm constant while varying capacity, enabling a controlled analysis of how structured lexicographic knowledge injection interacts with model size.

[3] https://huggingface.co/cjvt/GaMS-2B-Instruct

[4] https://huggingface.co/cjvt/GaMS-9B-Instruct

4.3. RAG Configuration

Knowledge injection was implemented via RAG pipeline over a structured lexicographic resource stored in JSONL format. Each record contains a sense identifier, idiom-level metadata, and a textual payload consisting of either (i) a definition (DEF), (ii) curated usage examples (EXS), or (iii) their combination (DEF+EXS), depending on the experimental condition.

Indexing phase. All JSONL records were embedded with `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`[5] (Reimers and Gurevych, 2019; Wang et al., 2020). The model produces 384-dimensional vectors, which we L2-normalized before indexing. Similarity search uses cosine similarity, implemented as inner-product search over normalized vectors in a FAISS `IndexFlatIP` index.

Retrieval phase. Retrieval is two-stage. We first attempt a deterministic hard match using idiom metadata (exact match to the target idiom). If no direct match is found, we fall back to dense retrieval over the FAISS index. Across all RAG conditions, retrieval parameters are fixed to top-k=1 with a maximum injected context length of 2000 characters.

Injection phase. The retrieved content is inserted into the prompt in a clearly demarcated reference section and treated as the sole external evidence used for the decision. The model is instructed to rely only on the reference text and the concordance sentence (baseline prompting omits the reference section). The retrieval pipeline and all RAG hyperparameters are identical for GaMS-2B-Instruct and GaMS-9B-Instruct, ensuring that differences across conditions stem from the type of injected lexicographic knowledge rather than retrieval configuration. The overall RAG pipeline is summarized in Figure 1.

4.4. Prompting Strategy

We use a single shared prompt template across the baseline and all RAG conditions to ensure strict comparability; the only differences across conditions are (i) whether a reference block is present (baseline: none; RAG: injected content) and (ii) a small condition-specific clarification of what constitutes semantic alignment (definitions vs. examples vs. both). The overall protocol is a structured, conservative two-step decision procedure.

First, the model assigns an alignment score $MAP\_SCORE \in \{0, 1, 2\}$ that reflects how clearly the usage in the concordance matches the available evidence (baseline: sentence context only; RAG: sentence + injected reference).

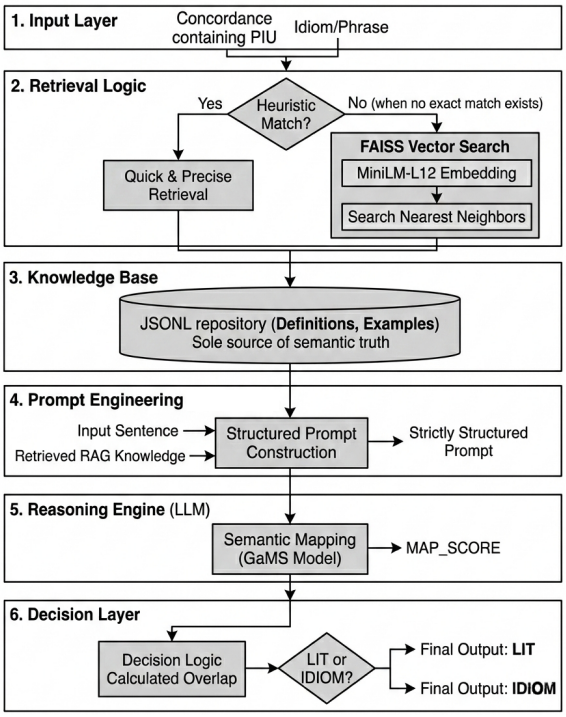


Figure 1: RAG protocol for Croatian literal–idiomatic disambiguation: deterministic metadata match with FAISS fallback, followed by prompt injection of structured lexicographic evidence.

Second, the final label is derived deterministically: only MAP_SCORE=2 yields IDIOMATIC, while $MAP\_SCORE \in \{0, 1\}$ defaults to LITERAL. This mapping enforces conservative grounding: the idiomatic label is allowed only when the model can confidently justify semantic correspondence.

To reduce format variability, all prompts require a fixed three-line output (decision, MAP_SCORE, and a short mapping/justification). We avoided open-ended expert-style prompts (e.g. requesting a linguistic analysis), which increased verbosity and structural variance across model sizes. Larger models also tended to produce more verbose outputs, reinforcing the need for a strictly structured format. The resulting structured prompting setup minimizes confounding effects due to prompt sensitivity and isolates the contribution of injected knowledge. A baseline prompt skeleton is shown below; full Croatian templates for the RAG variants are available online.[6]

```
PROMPT SKELETON
TASK: Decide LITERAL vs IDIOMATIC
for <PIE> in <SENTENCE>.
STEP 1: MAP_SCORE in {0,1,2}
        (semantic alignment)
   - 0: no evidence
   - 1: weak/uncertain evidence
```

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

[6] https://github.com/sbeliga/CroPIEs-1k

```
   - 2: clear evidence
   RESTRICTIONS:
     - sentence(+reference) only;
     - no external knowledge;
     - unsure => MAP_SCORE<2.
STEP 2: If MAP_SCORE=2 -> IDIOMATIC;
        Else -> LITERAL.
[REFERENCE BLOCK - RAG only]
<< injected  Def/Exs/Def+Exs >>
OUTPUT (3 lines):
DECISION; MAP_SCORE; JUSTIFICATION
```

4.5. Inference Configuration

All experiments use controlled decoding to attribute differences to knowledge injection rather than generation variability. We apply greedy decoding (`do_sample=False`, `num_beams=1`) with `repetition_penalty=1.0`, `use_cache=False`, and `max_new_tokens=60`, which is sufficient for the required short, structured outputs. We run mixed-precision inference (FP16) without quantization; a fixed random seed (1234) and deterministic execution are used where supported.

To avoid truncation effects, inputs are tokenized without truncation and a 300-token safety margin is reserved within the context window (fail-fast if exceeded), ensuring the full concordance context is preserved. When available, prompts are rendered with the tokenizer chat template (`apply_chat_template` with `add_generation_prompt=True`) for consistent formatting across models and conditions.

5. Experiments and Results

Table 3 summarizes PIE-level macro-F1 (mean±SD over 10 Croatian PIEs) for GaMS-2B-Ins. and GaMS-9B-Ins. under the baseline and three RAG knowledge-injection variants. Without external knowledge, both models perform modestly, with the larger model outperforming the smaller one (0.4377 vs. 0.3357). RAG improves performance for both models, with the strongest gains consistently obtained by **Def+Exs** (GaMS-2B-Ins.: 0.4584, $\Delta = +0.1227$; GaMS-9B-Ins.: 0.6211, $\Delta = +0.1834$). Definitional knowledge (DEF) yields larger average gains than example-only injection (EXS), suggesting that explicit semantic descriptions provide more stable disambiguation cues than contextual similarity alone. Overall, these results show that a small, manually curated phraseological resource can substantially improve Croatian literal–idiomatic disambiguation in a low-resource setting. Gains are observed even for the smaller model, suggesting that structured lexicographic evidence can partially compensate for limited internal representations in low-resource conditions. The relatively large SD values further indicate substantial heterogeneity across PIEs, motivating an idiom-level analysis.

To test whether gains generalize across PIEs, we use a one-sided exact Wilcoxon signed-rank test on idiom-level differences ($\Delta = F1_{\text{RAG}} - F1_{\text{Baseline}}$, $n = 10$ PIEs) to assess whether the median improvements (Δ) exceeds zero. Holm–Bonferroni correction is applied for multiple comparisons (Table 4). For GaMS-2B-Ins., only **Def+Exs** shows a statistically reliable improvement over the baseline ($p_{\text{Holm}} = 0.0391$, $r = 0.725$). For GaMS-9B-Ins., both DEF ($p_{\text{Holm}} = 0.0146$, $r = 0.822$) and **Def+Exs** ($p_{\text{Holm}} = 0.0059$, $r = 0.886$) are significant, while EXS is not. The effect sizes are large, indicating practically meaningful improvements in addition to statistical reliability.

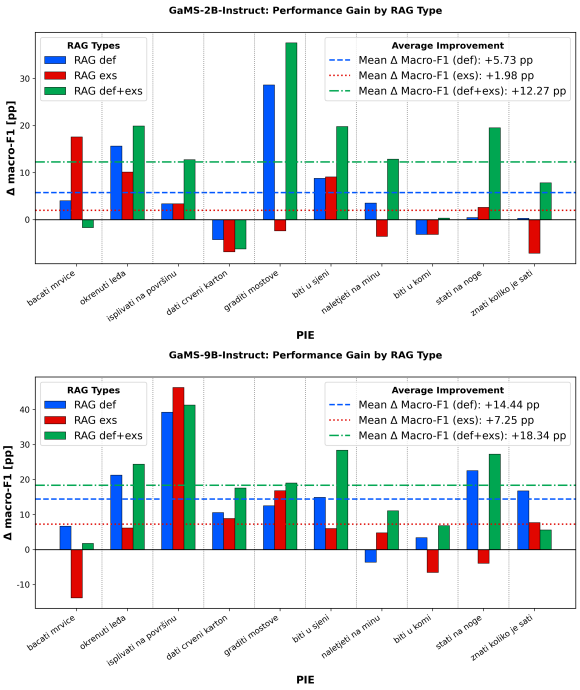


Figure 2: Idiom-level change in macro-F1 (RAG–baseline; percentage points) for 10 Croatian PIEs. Bars show Δ macro-F1 for three knowledge variants (DEF, EXS, DEF+EXS); horizontal lines indicate the mean Δ across PIEs for each variant (within each model).

Figure 2 breaks down these effects by PIE and contrasts the three RAG variants (bars show Δ in percentage points for interpretability; dashed lines denote mean Δ per variant within each model). Improvements are not uniform across expressions, but **Def+Exs** yields the most stable and frequently positive pattern across PIEs, suggesting that combining definitional and contextual evidence provides complementary signals for disambiguation. Notably, for GaMS-9B-Ins. the **Def+Exs** variant does not yield negative changes for any PIE, and over-

Setting	GaMS-2B-Ins.		GaMS-9B-Ins.	
	Macro-F1 (M ± SD)	Δ	Macro-F1 (M ± SD)	Δ
Baseline	0.3357 ± 0.0348	–	0.4377 ± 0.0746	–
RAG (Def)	0.3930 ± 0.1064	+0.0573	0.5821 ± 0.1201	+0.1444
RAG (Exs)	0.3555 ± 0.0530	+0.0198	0.5102 ± 0.1451	+0.0725
RAG (Def+Exs)	**0.4584 ± 0.1424**	**+0.1227**	**0.6211 ± 0.1278**	**+0.1834**

Table 3: PIE-level macro-F1 (mean ± sample SD over 10 Croatian PIEs) for GaMS-2B-Ins. and GaMS-9B-Ins. across the baseline and RAG variants. Δ denotes the mean absolute difference from the baseline, computed on the [0,1] scale.

Model	Config	p_{Holm}	r	Sig.
GaMS-2B-Ins.	Def	0.1260	0.564	–
GaMS-2B-Ins.	Exs	0.3125	0.177	–
GaMS-2B-Ins.	Def+Exs	0.0391	0.725	*
GaMS-9B-Ins.	Def	0.0146	0.822	*
GaMS-9B-Ins.	Exs	0.1934	0.435	–
GaMS-9B-Ins.	Def+Exs	0.0059	0.886	*

Table 4: Wilcoxon signed-rank tests over idiom-level macro-F1 scores ($n = 10$), comparing RAG configurations against the baseline. Holm–Bonferroni correction was applied. Effect size $r = Z/\sqrt{n}$. * indicates $p_{\text{Holm}} < 0.05$.

No.	PIE	Δ 2B	Δ 9B
1	*bacati mrvice*	-0.0168	0.0179
2	*biti u komi*	0.0034	0.0686
3	*biti u sjeni*	0.1980	0.2840
4	*dati crveni karton*	-0.0623	0.1759
5	*graditi mostove*	0.3757	0.1904
6	*isplivati na površinu*	0.1275	**0.4129**
7	*naletjeti na minu*	0.1285	0.1110
8	*okrenuti leđa*	**0.1992**	0.2442
9	*stati na noge*	0.1954	0.2726
10	*znati koliko je sati*	0.0784	0.0561

Table 5: Per-PIE macro-F1 improvement (Δ) of RAG (DEF+EXS) over the baseline for GaMS-2B-Ins. and GaMS-9B-Ins. (computed on the [0,1] scale).

all the 9B model exhibits fewer degradations than 2B across variants. In contrast, Exs is the least stable variant and occasionally produces negative changes, consistent with the intuition that example-only grounding can introduce noise when semantic alignment is weak. The largest gains tend to occur for more semantically opaque PIEs (e.g., *graditi mostove*, *isplivati na površinu*), supporting the usefulness of structured lexicographic knowledge for non-literal interpretation in a low-resource setting.

Table 5 reports per-PIE macro-F1 improvements for the best-performing configuration (DEF+EXS) relative to the baseline. Gains are not uniform across expressions. For GaMS-2B-Ins., DEF+EXS improves performance on 8/10 PIEs, with small drops limited to two cases (*bacati mrvice* and *dati crveni karton*); the largest gains are observed for *graditi mostove* and *okrenuti leđa*. GaMS-9B-Ins. shows more consistent behavior, improving on all 10/10 PIEs and achieving particularly strong gains for *isplivati na površinu* and *biti u sjeni*. Overall, these per-PIE results corroborate the macro-level trends in Table 3 and the distributional patterns in Figure 2, indicating broadly distributed gains rather than isolated outliers.

Table 6 reports idiom-averaged per-class performance, revealing pronounced prediction biases in the baseline condition. GaMS-2B-Ins. exhibits a strong LITERAL bias, achieving near-perfect LITERAL recall (0.9904) while almost completely failing to detect IDIOMATIC usage (R=0.0180, F1=0.0336). In contrast, GaMS-9B-Ins. shows the opposite tendency, strongly favoring IDIOMATIC predictions (R=0.9863) at the expense of LITERAL detection (R=0.1007). These opposing biases highlight that model scale alone does not guarantee balanced literal–idiomatic decisions.

Lexicographic knowledge injection reshapes these class-level behaviors. For GaMS-2B-Ins., all RAG configurations substantially increase IDIOMATIC recall (up to 0.9647 in EXS), confirming that injected evidence helps the smaller model recognize non-literal meaning. However, EXS overgeneralizes toward IDIOMATIC predictions, collapsing LITERAL recall to 0.0268. The combined DEF+EXS variant yields a more balanced trade-off, improving IDIOMATIC F1 (0.5507) while retaining moderate LITERAL performance.

For GaMS-9B-Ins., RAG primarily improves LITERAL recall (from 0.1007 to 0.6011 in DEF+EXS) while maintaining competitive IDIOMATIC performance. Across models, definitional grounding (DEF and DEF+EXS) has a stabilizing effect, whereas example-only injection (EXS) can shift the decision boundary and induce class overprediction in a model-dependent manner. Overall, these results show that structured lexicographic resources improve not only macro-F1 but also the balance between LITERAL and IDIOMATIC predictions by mitigating baseline bias. This is consistent with the conservative MAP_SCORE protocol, where stronger evidence in the reference block can systematically shift the model toward (or away from) the idiomatic decision.

Model	Config	Literal			Idiomatic		
		P	R	F_1	P	R	F_1
GaMS-2B-Ins.	Baseline	0.4750	0.9904	0.6377	0.3000	0.0180	0.0336
	RAG (Def)	0.3046	0.2231	0.2147	0.5024	0.7567	0.5713
	RAG (Exs)	0.2650	0.0268	0.0385	**0.5243**	**0.9647**	**0.6724**
	RAG (Def+Exs)	0.5053	0.3829	0.3660	0.4953	0.6728	0.5507
GaMS-9B-Ins.	Baseline	0.8079	0.1007	0.1741	0.5492	0.9863	0.7013
	RAG (Def)	0.6039	0.5486	0.5208	0.6715	0.6957	0.6434
	RAG (Exs)	**0.5228**	**0.8442**	**0.6337**	0.7126	0.3029	0.3867
	RAG (Def+Exs)	0.6401	0.6011	0.5760	0.6957	0.7026	0.6661

Table 6: Mean per-class performance across idioms. Precision (P), Recall (R), and F1 scores are computed separately for the Literal and Idiomatic classes and averaged over the 10 target idioms for each model and configuration.

6. Discussion

Our results show that inference-time injection of structured lexicographic knowledge can substantially improve Croatian PIE literal–idiomatic disambiguation for locally runnable, open-weight decoder-only LLMs without fine-tuning. Across both GaMS models, RAG yields macro-level gains, with DEF+EXS producing the most reliable improvements and the strongest statistical evidence. This pattern supports the view that definitions and curated usage examples provide complementary signals: definitions act as stable semantic anchors, while examples contribute prototypical contextual cues for matching concordance usage.

At the same time, gains are not uniform across expressions. PIE-level analyses indicate that some expressions benefit strongly, whereas a small subset shows marginal improvements or occasional degradations, highlighting sensitivity to contextual fit and the structure of retrieved evidence. Importantly, the per-class breakdown shows that lexicographic grounding does not merely increase aggregate scores: it can reshape decision behavior by mitigating strong baseline prediction biases (literal-biased GaMS-2B vs. idiomatic-biased GaMS-9B). Example-only grounding is also less stable and can induce class overprediction in a model-dependent way, reinforcing the stabilizing role of definitional evidence.

A limitation of the current setup is that the target PIE can appear in morphologically and syntactically varied realizations within concordances (e.g., inflectional changes, word-order variation, or the insertion of additional words (e.g., modifiers or clitics) within the expression, and occasionally multiple times in the same instance. While our evaluation assumes the target expression is present, further error analysis should disentangle potential failures in (i) identifying the intended target span under such variation and (ii) performing literal–idiomatic disambiguation once the target is identified. Additional confounds arise when semantically related expressions occur nearby: for example, our study targets *dati crveni karton* (*to give a red card*), but some concordances also contain *dati žuti karton* (*to give a yellow card*), which may prime an idiomatic interpretation even though it is not the target of classification. Moreover, a small portion of concordances contains dialectal or informal Croatian; since GaMS is not explicitly trained or optimized for Croatian dialectal varieties or slang (and such data are likely underrepresented), these inputs may reduce both retrieval fit and model comprehension. Finally, structured outputs with brief justification appear more stable than single-label answers, motivating deeper analysis of how explanation requirements interact with grounding and decision reliability.

Despite these limitations, the results consistently indicate that small, expert-curated lexicographic resources provide effective *small data* grounding for non-literal interpretation in low-resource settings. In particular, several PIEs benefit robustly from injected evidence across models and knowledge variants, whereas a few remain challenging even under DEF+EXS, suggesting that targeted analysis of dictionary senses, example selection, and contextual ambiguity could further improve grounding quality.

Overall, these findings align with a small-data, neurosymbolic perspective: compact, expert-curated lexicographic resources can compensate for representational gaps in locally adapted LLMs and provide controlled semantic evidence for non-literal interpretation in low-resource settings. The present findings are derived from a controlled dataset with a limited number of phraseological units. Future work will extend the evaluation to broader PIE inventories and retrieval strategies, and to a larger and more diverse set of idioms, allowing us to evaluate the approach under more heterogeneous conditions.

7. Conclusion

This paper presented a controlled study of Croatian PIE literal–idiomatic disambiguation, evaluat-

ing inference-time injection of structured lexicographic knowledge via RAG on locally runnable, open-weight decoder-only LLMs without fine-tuning. The study isolates the contribution of three knowledge variants: definitions (DEF), curated usage examples (EXS), and their combination (DEF+EXS), under a unified prompting and decoding protocol across two model sizes.

Across models and expressions, lexicographic grounding improves macro-F1, with DEF+EXS yielding the most reliable gains and the strongest statistical evidence. Definitional evidence is more stable than examples alone, while example-only grounding can be less consistent and may shift class-level behavior in a model-dependent way. Beyond aggregate scores, injected knowledge mitigates strong baseline prediction biases and improves the balance between LITERAL and IDIOMATIC decisions.

For reproducibility, we release the CroPIEs-1k expert-annotated concordance dataset and the prompt templates used in this study. Overall, the findings highlight that *small data* in the form of compact, expert-curated structured lexicographic knowledge can provide effective grounding for non-literal language understanding in low-resource settings without fine-tuning.

8. Acknowledgements

This research was supported by the project Hybrid AI Approaches to Natural Language Processing and Knowledge Generation – HyAI (uniri-iz-25-215), funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the author and do not necessarily reflect the official stance of the European Union or the European Commission. Neither the European Union nor the European Commission can be held accountable for them.

9. Bibliographical References

Hugo Abonizio, Thales Almeida, Roberto Lotufo, and Rodrigo Nogueira. 2025. Comparing knowledge injection methods for llms in a low-resource regime. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 819–830, Porto Alegre, RS, Brasil. SBC.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton.

Slobodan Beliga and Ivana Filipović Petrović. 2024. Large language models supporting lexicography: Conceptual organization of croatian idioms. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 23–46, Ljubljana. Institute of Contemporary History.

Slobodan Beliga and Ivana Filipović Petrović. 2025. Ai- and corpus-based strategies for identifying phraseme constructions: A pilot study on croatian repetitive constructions. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography*, pages 95–115, Brno. Lexical Computing CZ s.r.o.

Kushagra Bhushan, Yatin Nandwani, Dinesh Khandelwal, Sonam Gupta, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. 2025. Systematic knowledge injection into large language models via diverse augmentation for domain-specific RAG. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5937–5958, Albuquerque, New Mexico. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Ivana Filipović Petrović and Slobodan Beliga. 2025. Can ai understand croatian idioms? assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*, 65(3):218–242.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229, Mexico City, Mexico. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. ACL.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Dario Vajda, Domen Vreš, and Marko Robnik Šikonja. 2025. Improving LLMs for machine translation using synthetic preference data. In *Proceedings of the 2nd LUHME Workshop*, pages 67–73, Bologna, Italy. UP - Universidade do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JTDH 2024)*, pages 485–511. Institute of Contemporary History, Ljubljana, Slovenia.

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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.

Tadej Škvorc and Marko Robnik-Šikonja. 2025. Solving word-sense disambiguation and word-sense induction with dictionary examples.

10. Language Resource References

CLARIN.SI. 2024. *CLASSLA-web.hr 2.0*. Croatian web corpus, accessible via NoSketch Engine.

Croatian Academy of Sciences and Arts. 2023. *Online Dictionary of Croatian Idioms (Frazeološki rječnik hrvatskoga jezika), Version 2*. Open-access digital phraseological dictionary.

Metaphor Identification in Spanish Oncological Discourse: The Role of Explicit Meaning in Low-Resource Settings

Lucía Pitarch, Jordi Bernad, Gemma Bel-Enguix
Universidad de Zaragoza, Universidad Nacional Autónoma de México
Zaragoza (Spain), Ciudad de México (México)
{lpitarch, jbernad}@unizar.es, gbele@iingen.unam.mx

Abstract

Metaphor identification remains challenging in specialized and low-resource domains, where large annotated datasets are unavailable and general-domain models often fail to transfer effectively. In this paper, we evaluate FLAVORS-AECC, a Spanish dataset of oncological discourse that provides transparent, instance-level annotations of basic meaning (BM) and contextual meaning (CM) following the Metaphor Identification Procedure (MIP). We test the state-of-the-art Contrast-WSD model under two splits: a random split and a lemma-based split to control for lexical memorization. We compare three configurations: (i) a control model with no meaning information, (ii) manually curated basic meanings, and (iii) first dictionary entry as an approximation of basic meaning. Results show that explicitly modeling meaning contrast substantially improves performance in low-resource settings (from below 0.30 to above 0.50 F1). However, contrary to expectations, manually annotated BM does not consistently outperform first dictionary entries, suggesting that definition length rather than theoretical fidelity may introduce noise. We also find that models perform best on cases with high annotator agreement and that verbs remain the most challenging part of speech. Overall, our findings highlight the importance of linguistically grounded modeling for metaphor detection in specialized domains.

Keywords: Metaphor annotation, Spanish resources, oncological discourse, figurative language, low-resource NLP

1. Introduction

Metaphors are a fundamental linguistic device in medical communication, enabling speakers to explain complex or abstract experiences through more concrete domains (Lakoff and Johnson, 1980). In oncology, metaphors such as *cancer as war* or *the body as a battlefield* are pervasive and shape how patients, clinicians, and families conceptualize illness, treatment, and prognosis (Semino et al., 2017). Despite their importance, computational analysis of medical metaphors remains underexplored, particularly for languages other than English.

Traditional research on medical metaphors has largely relied on qualitative linguistic and discourse-analytic methods (Liu et al., 2024). While this work provides rich theoretical insights, it is labor-intensive, difficult to scale, and poorly suited for real-time processing of large volumes of patient narratives. Moreover, most existing annotated resources are English-centric, leaving a substantial gap for Spanish medical discourse.

In parallel, recent advances in automated metaphor identification have achieved strong performance on large general-domain datasets such as the VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017). However, these systems struggle in low-resource, domain-specific settings where training data is limited, and language use diverges from everyday contexts.

Metaphor is a complex linguistic phenomenon approached from multiple theoretical perspectives, yielding different annotation methodologies, from explicit and deliberate metaphor frameworks (Dipper et al., 2024) to dynamic approaches such as Cameron's (Cameron, 2007). Any computational system's operationalization of metaphor is therefore inevitably tied to the annotation guidelines of its gold standard. The most widely used benchmark dataset for automated metaphor identification, the VUA corpus, was annotated following the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007), and so have many of the systems trained and evaluated on it. MIP defines metaphoricity through a contrast between contextual meaning (CM): the sense a word takes in context, and basic meaning (BM): the most concrete, physical, and imaginable sense of a word. This contrast is what defines metaphoricity. We adopt MIP as our working framework because it underlies our gold standard, is among the most replicable and cross-linguistically documented procedures, and enables comparability across studies. A central issue concerns how BM is operationalized computationally. Many systems approximate it using the first dictionary entry or static embeddings (Choi et al., 2021), diverging from MIP's requirement that BM reflect the most concrete sense rather than the most frequent one. We acknowledge that word sense inventories are imperfect proxies, particularly for creative metaphors. Nevertheless, in low-resource, specialized domains, we hypothesize that the manual selection and annotation of BM–CM contrasts

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 31–39
11 May 2026.

provide better guidance for the model's predictions.

To test this hypothesis, we evaluate the FLAVORS-AECC dataset (Pitarch et al., 2026) using the state-of-the-art Contrast-WSD model (Elzohbi and Zhao, 2024). FLAVORS-AECC is the first Spanish dataset that provides transparent annotations of both basic at the instance level. It also includes two evaluation splits: a random split and a lemma-based split designed to reduce lexical memorization.

Our contributions are:

- The first domain-specific evaluation of automated metaphor identification in Spanish oncological discourse.
- A comparison between theoretically aligned models to annotation procedures.
- Analysis of differences between manually annotated BM and dictionary-based approximations.
- An analysis of how annotator agreement and part-of-speech affect model performance.

2. Related Work

The Metaphor Identification Procedure (MIP), introduced by the PRAGGLEJAZ Group in 2007 (Praglejaz Group, 2007), was a major step toward standardizing metaphor annotation at a time when labeling practices were highly heterogeneous. MIP defines a four-step procedure: (1) segmenting the text into lexical units, (2) identifying the basic meaning of each unit (the most concrete, imaginable, and tangible sense listed in a dictionary), (3) determining its contextual meaning in the given sentence, and (4) marking the unit as metaphorical if the two meanings contrast.

While intentionally minimalist to allow adaptation, MIP's flexibility has also led to substantial variation and subjectivity in annotation practices. Challenges arise in defining lexical units (De Backer et al., 2023), interpreting contextual meaning, and operationalizing basic meaning (Maudslay and Teufel, 2022). Although subjectivity is unavoidable, a key issue is the lack of transparency in annotation decisions. Existing datasets often describe their segmentation criteria (Krennmayr and Steen, 2017; Sanchez-Bayona and Agerri, 2022; Sánchez-Montero et al., 2025), but rarely make explicit, for each instance, how basic and contextual meanings are interpreted. To the best of our knowledge, FLAVORS-AECC is the only dataset that explicitly encodes these meanings via WordNet synset annotations, ensuring transparency and consistency.

Regarding automated metaphor identification, most computational approaches to metaphor identification are inspired by MIP and aim to operationalize basic meaning within neural architectures. MelBERT (Choi et al., 2021) has been particularly influential, inspiring a range of subsequent models (Elzohbi and Zhao, 2024; Babieno et al., 2022; Li et al., 2023). Many of these systems attempt to operationalize basic meaning (BM) within neural architectures. Some approaches assume BM is captured by static or decontextualized embeddings (Song et al., 2021; Choi et al., 2021), while others approximate BM using the first dictionary definition of a word (Su et al., 2021; Babieno et al., 2022). However, both strategies diverge from MIP, as frequency-based meanings or first dictionary senses do not necessarily correspond to the most concrete or physical sense.

Building on these insights, our work advocates for explicit, manually curated basic meaning annotations aligned with MIP and directly provided to the model enhancing transparency, interpretability, and theoretical coherence.

3. Experimental Setup

We conduct our experiments on a filtered subset of the FLAVORS-AECC dataset (Pitarch et al., 2026). To ensure compatibility with the VUA format and the Contrast-WSD model, we retain only single-word metaphor annotations and restrict the data to verbs, nouns, and adjectives. The resulting dataset contains 5,239 instances with annotated basic and contextual meanings, of which 18% are labeled as metaphorical by at least one annotator. The overall inter-annotator agreement reported for the original dataset is 0.49 F1[1]. The part-of-speech distribution is 64% verbs, 24% nouns, and 12% adjectives. A sample of the dataset is shown in Table 1.

We evaluate all experiments under two data splitting strategies: a random and a lemma-based split. In the random split the data is divided into train and test sets using an 80/20 ratio while preserving the proportion of metaphorical and literal instances. The lemma-based split enforces that target lemmas do not overlap between train and test sets, thereby reducing lexical memorization effects.

We use Contrast-WSD (Elzohbi and Zhao, 2024), a state-of-the-art metaphor identification model inspired by the Metaphor Identification Procedure (MIP). Figure 1 presents an overview of the architecture. The model takes as input a sentence, a target word within the sentence, a definition representing the word's basic meaning, and a definition corresponding to the word's contextual meaning.

A RoBERTa model is used to obtain embeddings: a) of the target word in the full sentence with spe-

[1] F1 score was chosen against the common Kappa score, as suggested by (Boguslav and Cohen, 2017) for flexible span annotations. More details on this choice in the dataset original paper (Pitarch et al., 2026)

ID	w_index	pos	lemma	sentence	agree	label	basic meaning (BM)	contextual meaning (CM)
6_s2_w1	1	N	hermano	Mi hermano con 30 años acaba de ser diagnosticado con un tumor en el pulmón.	2	0	a male with the same parents as someone else	close friend who accompanies his buddies in their activities
6_s3_w0	0	V	tener	Tienen sospecha de que sea maligno y estamos esperando cita con Cirugía	2	0	have or hold in one's hands or grip	have or possess, either in a concrete or an abstract sense
6_s3_w5	5	ADJ	maligno	Tienen sospecha de que sea maligno y estamos esperando cita con Cirugía	2	1	maligno, malvado, maléfico, malévolo	canceroso, maligno

Table 1: AECC-FLAVORS in VUAM format processed sample. Label=1 means metaphoric example, Label=0 means non-metaphoric instance. In this case only *maligno* is annotated as metaphoric.

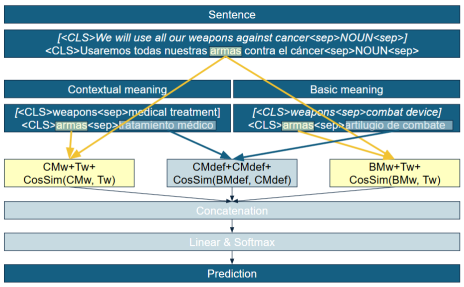


Figure 1: ContrasWSD arquitecture schema. In the original model (Elzohbi and Zhao, 2024) BM definition is the first dictionary entry. '+' signs represent concatenation.

cial tokens marking the target word and POS; b) of the target word in the contextual definition and of the contextual definition; c) of the target word in the basic meaning definition and of the basic meaning. The contextual and basic meaning embeddings are concatenated jointly with its cosine similarity. The same process (concatenate jointly with the cosine similarity) is performed with the target word embeddings in the full sentence and in the contextual meaning definition, and with the target word embeddings in the full sentence and in the basic meaning definition. These three embeddings are concatenated and fed into a final classification layer that predicts whether the target word is used metaphorically or literally.

While we use the same models as Elzohbi and Zhao (2024), we differ in our augmentation of the data. Where they use an additional external model for word sense disambiguation, we directly use Lesk (Lesk, 1986) algorithm. And secondly, while they only use the first Wikitionary dictionary entry as basic meaning, we compare the manual selection of basic meaning with the first wordnet dictionary entry. All definitions are extracted from WordNet using the corresponding synsets and are provided in Spanish when available, and in English otherwise.

We evaluate three experimental configurations:

1. **Control**: only the sentence and target word are provided, without any definitions.
2. **Manual Basic Meaning**: the basic meaning is manually annotated using the most appropriate WordNet synset.
3. **First Dictionary Entry**: the basic meaning is approximated using the first available dictionary definition.

We explore multiple hyperparameter configurations (see Appendix for full details) and select the best-performing setup with class weight[2] $= 5$, learning rate $= 1 \times 10^{-5}$, batch size $= 16$, warm-up epochs $= 2$, and total epochs $= 10$. Since definitions may be provided in both Spanish and English, we use XLM-RoBERTa-base (Conneau et al., 2019) as the encoder. Each experiment is run five times with different random seeds, and we report mean performance along with 95% confidence intervals.

4. Results

Table 2 presents the main quantitative results of our experiments for the three modeling configurations (Control, Manual Basic Meaning, and First Dictionary Entry) under both evaluation splits (random by label and lemma-based). Scores correspond to the mean over five runs, using the best-performing hyperparameter configuration.

trainingSplit	BMAugmentation	Precission	Recall	F1
lemma	Control	0.311	0.243	0.273
lemma	Control	0.308	0.280	0.292
lemma	ManualBM	0.259	0.507	0.342
lemma	ManualBM	0.270	0.526	0.357
lemma	1stDicEntry	0.277	0.508	0.360
lemma	1stDicEntry	0.296	0.541	0.383
random	1stDicEntry	0.414	0.617	0.495
random	ManualBM	0.426	0.607	0.501
random	1stDicEntry	0.455	0.568	0.505
random	ManualBM	0.462	0.580	0.514

Table 2: Results ordered by F1 (lowest to highest) to asses the best configuration of three model augmentation settings (manual BM selection, 1st dictionary entry as BM, and control: no added basic nor contextual meaning information). Both random and lemma splits are displayed.

[2]Class weight is a parameter in Elzohbi and Zhao (2024) arquitecture which weights more metaphoric instances than non metaphoric ones to balance the difference between non metaphoric sentences and metaphoric ones.

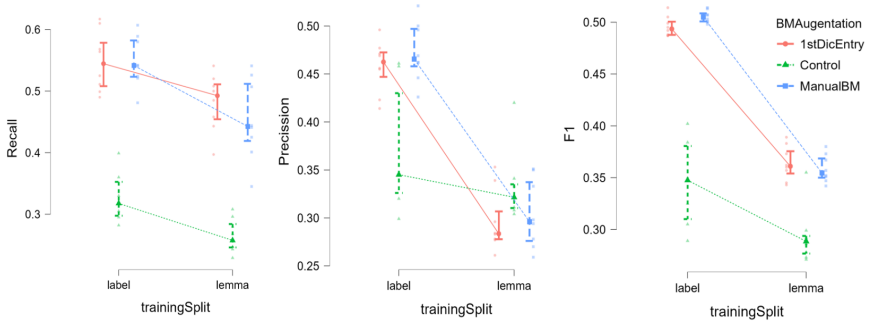


Figure 2: Performance metrics across different configurations and splits

Overall, the results reveal three central patterns that align with our research questions. First, models that explicitly incorporate information about basic and contextual meaning substantially outperform the control condition that relies solely on sentence context and the target word. Under the lemma-based split, the control model achieves only 0.27 F1, whereas the two meaning-aware configurations reach up to 0.38 F1. The effect is even more pronounced under the random split, where performance increases from below 0.30 to above 0.50 F1. This confirms our main hypothesis: in low-resource, domain-specific settings, providing linguistically grounded meaning representations is more beneficial than relying on data quantity alone.

Second, contrary to our expectations, manually annotated basic meanings do not consistently outperform the first dictionary entry approximation. Under the random split, Manual BM yields the best result (0.51 F1), but the difference with the dictionary-based approach is marginal (0.50 F1). More strikingly, under the lemma-based split, the first dictionary entry slightly outperforms the manual annotations. It is worth noting that the competitive performance of the first-sense heuristic is itself a well-established finding in Word Sense Disambiguation (McCarthy et al., 2007), where it has long served as a strong baseline. Our surprise, however, stems not from its general effectiveness but from its performance in this specific setting: MIP explicitly instructs annotators not to rely on the first dictionary entry when identifying basic meaning, emphasizing instead the most concrete, physical, and imaginable sense, which may differ from the most frequent one. The marginal gains of manual BM annotations thus suggest that theoretical fidelity to MIP's definition does not automatically translate into empirical gains in this context. One plausible explanation is that manual definitions tend to be longer and more detailed, potentially introducing noise into the model, whereas shorter dictionary definitions provide a more stable and compact semantic signal, one that, despite diverging from MIP's theoretical intent, proves empirically competitive.

Third, the best overall performance (0.51 F1) is comparable to the reported inter-annotator agreement of the dataset (0.49 F1). This suggests that the task is inherently difficult and that further improvements may require richer linguistic modeling rather than more data alone. Matching human agreement is expected in this setting; exceeding it would raise concerns about overfitting or unintended annotation leakage.

Figure 2 visualizes the relative gains across conditions, highlighting that the primary performance jump occurs when any form of basic and contextual meaning is introduced. This pattern contrasts with findings in the original Contrast-WSD study, where meaning information yielded only modest improvements on larger, general-domain datasets (0.74 vs. 0.72 F1). Our results indicate that explicit semantic modeling becomes crucial precisely in the most challenging scenarios: small datasets, specialized domains, and less-resourced languages.

4.1. Error Analysis

To better understand model behavior, we conducted a targeted error analysis focusing on annotator agreement and part-of-speech (POS) effects.

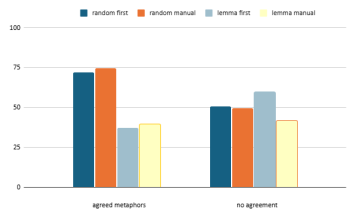


Figure 3: Error analysis in relation to annotator agreement.

Figure 3 shows performance stratified by in-

stances with full annotator agreement versus disagreement. As expected, cases where both annotators labeled a token as metaphorical are significantly easier for the model to predict. This effect is especially pronounced under the lemma-based split, suggesting that when lexical memorization is minimized, the model relies more on genuinely prototypical metaphoric patterns. Conversely, instances with annotator disagreement remain particularly challenging, indicating that these cases are ambiguous even for humans.

Figure 4 reports performance by part of speech. Adjectives emerge as the most robust category, showing relatively stable performance across splits and modeling conditions. Verbs, in contrast, are consistently the most difficult to classify. We attribute this to their greater semantic complexity: verbs encode events, relations, and argument structures that are not fully captured by simple definition-based representations. We therefore propose that future work should incorporate additional linguistic features for verbs, such as valence, semantic roles, or event structure, which may help the model better distinguish literal from metaphorical uses.

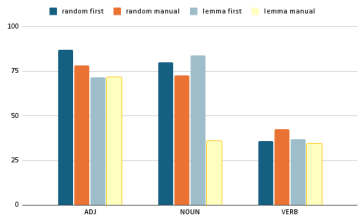


Figure 4: Error analysis in relation to POS

5. Conclusions

Taken together, these results demonstrate that (1) meaning-aware architectures are crucial in low-resource metaphor identification, (2) simpler approximations of basic meaning can be competitive with manual annotations, and (3) substantial gains will likely require more linguistically informed modeling rather than larger datasets alone.

Beyond the technical contributions, these findings carry direct implications for the medical domain that originally motivated this work. The Spanish oncological forum of the Asociación Española Contra el Cáncer was manually annotated at considerable cost — over 88,000 words, six annotators, two years of effort, and consultation across multiple disciplinary experts, yielding the gold-standard resource used throughout this study. The methods evaluated here open a path toward extending this coverage to the full forum, comprising over five million words, without incurring equivalent annotation costs. Such large-scale metaphor identification would enable the study of metaphor as a dynamic, longitudinal process: how figurative language shifts across disease phases, how patients adapt their expression over time, and what linguistic patterns may signal changes in emotional state or coping strategies. This connects directly to the agenda outlined in (Pitarch and Bel-Enguix, 2026), where metaphor is framed not as a static lexical phenomenon but as an evolving communicative resource shaped by the patient's trajectory. We therefore see the present work not as a closed contribution, but as a methodological stepping stone toward a richer, data-driven understanding of illness narratives in online health communities.

6. Acknowledgements

This paper has been supported by PAPIIT project IG-400325, by the I+D+i projects PID2024-159530OB-I00 (funded by MCIN/AEI/10.13039/501100011033), the EU research and innovation program HORIZON Europe in the "4D PICTURE" project under grant agreement 101057332 and UZ2024-IyA-02 (funded by Univ. Zaragoza), and by DGA Government predoctoral fellowship.

7. Bibliographical References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4).

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 298–302. IOS Press.

Lynne Cameron. 2007. The affective discourse dynamics of metaphor clustering. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, (53):041–062.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Laurence De Backer, Renata Enghels, and Patrick Goethals. 2023. Metaphor analysis meets lexical strings: finetuning the metaphor identification procedure for quantitative semantic analyses. *Frontiers in Psychology*, 14:1214699.

Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-My Nguyen. 2024. Guidelines for the annotation of deliberate linguistic metaphor. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Mohamad Elzohbi and Richard Zhao. 2024. ContrastWSD: Enhancing metaphor detection with word sense disambiguation following the metaphor identification procedure. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3907–3915, Torino, Italia. ELRA and ICCL.

Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Yufeng Liu, Elena Semino, Judith Rietjens, and Sheila Payne. 2024. Cancer experience in metaphors: patients, carers, professionals, students – a scoping review. *BMJ Supportive & Palliative Care*, 14(e3):e2366–e2376.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical polysemy detection: Conventional metaphor meets word sense disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 65–77.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

Lucía Pitarch and Gemma Bel-Enguix. 2026. Modeling metaphor evolution on cancer online narratives. In *Proceedings of the 16th International Conference on the Evolution of Language (EVOLANG XVI)*.

Lucía Pitarch, Jordi Bernad, Sergio-Luis Ojeda-Trueba, Alec Sánchez-Montero, Max Ionov, Emma Anglés-Herrero, Ángel Óscar Corona Beomont, and Gemma Bel-Enguix. 2026. Medical-flavors-aecc: Spanish oncological metaphors dataset. In Press. Accepted at CL4H@LREC 2026.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. 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)*, pages 4240–4251, Online. Association for Computational Linguistics.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1280–1287, Online. Association for Computational Linguistics.

Appendix

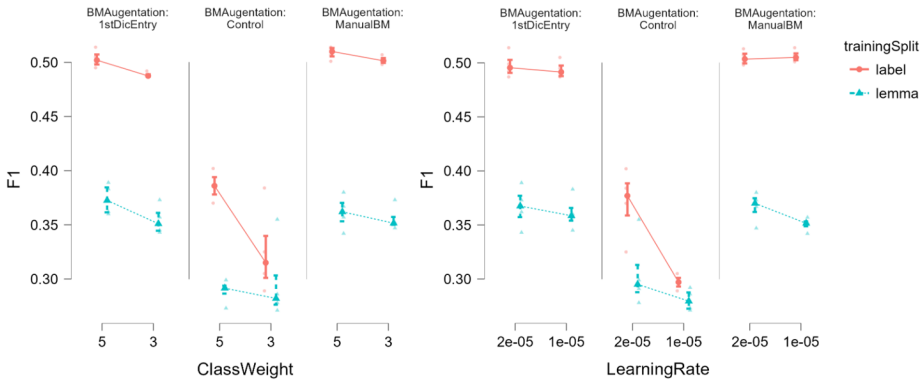


Figure 5: Learning Rate and Class Weight Plot

trainingSplit	BMAugentation	ClassWeight	LearningRate	Precission	PrecCI	Recall	RecCI	F1	F1CI
lemma	Control	3	1,00E-05	0.333	(0.290–0.375)	0.229	(0.195–0.263)	0.271	(0.236–0.305)
lemma	Control	5	1,00E-05	0.311	(0.270–0.352)	0.243	(0.208–0.278)	0.273	(0.240–0.307)
lemma	Control	3	2,00E-05	0.322	(0.283–0.361)	0.247	(0.215–0.279)	0.278	(0.245–0.312)
lemma	Control	3	1,00E-05	0.341	(0.300–0.386)	0.247	(0.215–0.279)	0.286	(0.250–0.322)
label	Control	3	1,00E-05	0.299	(0.245–0.360)	0.282	(0.230–0.339)	0.289	(0.239–0.345)
lemma	Control	5	2,00E-05	0.321	(0.283–0.364)	0.268	(0.234–0.304)	0.291	(0.257–0.327)
lemma	Control	5	1,00E-05	0.308	(0.272–0.345)	0.280	(0.245–0.313)	0.292	(0.259–0.325)
lemma	Control	5	2,00E-05	0.304	(0.266–0.340)	0.296	(0.260–0.333)	0.299	(0.266–0.332)
label	Control	3	1,00E-05	0.320	(0.260–0.376)	0.295	(0.244–0.348)	0.305	(0.249–0.353)
label	Control	3	2,00E-05	0.346	(0.303–0.391)	0.305	(0.268–0.345)	0.325	(0.288–0.361)
lemma	ManualBM	5	1,00E-05	0.259	(0.237–0.284)	0.507	(0.467–0.549)	0.342	(0.316–0.368)
lemma	1stDicEntry	3	2,00E-05	0.261	(0.237–0.289)	0.500	(0.462–0.539)	0.343	(0.314–0.370)
lemma	1stDicEntry	3	1,00E-05	0.284	(0.257–0.312)	0.443	(0.403–0.481)	0.345	(0.314–0.376)
lemma	ManualBM	3	2,00E-05	0.350	(0.314–0.386)	0.345	(0.309–0.381)	0.347	(0.314–0.383)
lemma	ManualBM	3	1,00E-05	0.298	(0.269–0.327)	0.425	(0.388–0.463)	0.351	(0.320–0.380)
lemma	ManualBM	3	1,00E-05	0.294	(0.266–0.325)	0.441	(0.403–0.481)	0.352	(0.323–0.383)
lemma	Control	3	2,00E-05	0.420	(0.323–0.526)	0.308	(0.228–0.390)	0.355	(0.272–0.435)
lemma	1stDicEntry	3	1,00E-05	0.283	(0.254–0.309)	0.485	(0.446–0.525)	0.357	(0.330–0.386)
lemma	ManualBM	5	1,00E-05	0.270	(0.245–0.296)	0.526	(0.484–0.565)	0.357	(0.329–0.384)
lemma	1stDicEntry	5	1,00E-05	0.277	(0.252–0.301)	0.508	(0.470–0.547)	0.360	(0.333–0.387)
lemma	1stDicEntry	5	2,00E-05	0.278	(0.249–0.304)	0.520	(0.481–0.556)	0.362	(0.335–0.388)
lemma	ManualBM	5	2,00E-05	0.278	(0.255–0.304)	0.541	(0.502–0.577)	0.367	(0.340–0.396)
label	Control	5	2,00E-05	0.344	(0.314–0.377)	0.399	(0.364–0.436)	0.370	(0.340–0.404)
lemma	1stDicEntry	3	2,00E-05	0.353	(0.318–0.387)	0.397	(0.358–0.433)	0.373	(0.342–0.405)
lemma	ManualBM	3	2,00E-05	0.351	(0.318–0.386)	0.401	(0.365–0.440)	0.373	(0.345–0.405)
lemma	ManualBM	5	2,00E-05	0.333	(0.300–0.367)	0.444	(0.407–0.482)	0.380	(0.349–0.412)
lemma	1stDicEntry	5	1,00E-05	0.296	(0.270–0.321)	0.541	(0.503–0.581)	0.383	(0.355–0.413)
label	Control	3	2,00E-05	0.461	(0.418–0.503)	0.330	(0.294–0.365)	0.384	(0.350–0.420)
lemma	1stDicEntry	5	2,00E-05	0.339	(0.305–0.371)	0.458	(0.420–0.495)	0.389	(0.356–0.422)
label	Control	5	2,00E-05	0.458	(0.416–0.500)	0.360	(0.327–0.394)	0.402	(0.369–0.438)
label	1stDicEntry	3	2,00E-05	0.456	(0.422–0.492)	0.525	(0.488–0.563)	0.487	(0.456–0.518)
label	1stDicEntry	3	1,00E-05	0.469	(0.435–0.503)	0.511	(0.474–0.550)	0.487	(0.455–0.521)
label	1stDicEntry	3	1,00E-05	0.477	(0.443–0.510)	0.499	(0.463–0.535)	0.488	(0.458–0.518)
label	1stDicEntry	3	2,00E-05	0.496	(0.458–0.530)	0.490	(0.452–0.528)	0.492	(0.460–0.526)
label	1stDicEntry	5	1,00E-05	0.414	(0.382–0.442)	0.617	(0.580–0.653)	0.495	(0.468–0.523)
label	ManualBM	3	2,00E-05	0.462	(0.428–0.497)	0.541	(0.504–0.580)	0.498	(0.470–0.530)
label	1stDicEntry	5	2,00E-05	0.423	(0.393–0.453)	0.610	(0.576–0.644)	0.499	(0.468–0.528)
label	ManualBM	3	2,00E-05	0.521	(0.482–0.557)	0.481	(0.443–0.521)	0.500	(0.468–0.537)
label	ManualBM	5	1,00E-05	0.426	(0.396–0.456)	0.607	(0.570–0.644)	0.501	(0.471–0.532)
label	ManualBM	3	1,00E-05	0.469	(0.435–0.502)	0.542	(0.503–0.580)	0.503	(0.471–0.535)
label	1stDicEntry	5	1,00E-05	0.455	(0.423–0.488)	0.568	(0.535–0.604)	0.505	(0.473–0.533)
label	ManualBM	3	1,00E-05	0.496	(0.458–0.533)	0.521	(0.482–0.559)	0.507	(0.474–0.538)
label	ManualBM	5	2,00E-05	0.446	(0.414–0.480)	0.589	(0.552–0.624)	0.507	(0.479–0.536)
label	ManualBM	5	2,00E-05	0.500	(0.464–0.536)	0.524	(0.486–0.562)	0.513	(0.484–0.542)
label	1stDicEntry	5	2,00E-05	0.471	(0.438–0.505)	0.564	(0.526–0.598)	0.514	(0.482–0.547)
label	ManualBM	5	1,00E-05	0.462	(0.429–0.493)	0.580	(0.543–0.617)	0.514	(0.485–0.545)

Table 3: Extended version of Table 2 including confidence intervals for each evaluation metric and the different hyperparameter configurations (class weight and learning rate). Results are ordered by F1 (lowest to highest) and display the three model augmentation settings (manual BM selection, 1st dictionary entry as BM, and control: no added basic nor contextual meaning information) under both random and lemma splits.

Exploring Detection of Complex, Non-Literal Expressions of Cultural Motifs

Ibrahim H. Alyami[1] **Mark A. Finlayson**[2]
[1]Najran University, King Abdul Aziz Rd, 66462, Najran, KSA
[2]Florida International University, 11200 SW 8th Street, Miami, FL, 33199, USA
ihalmerdef@nu.edu.sa, markaf@fiu.edu

Abstract

Motifs are non-commonplace, recurring narrative elements, often found originally in folk stories and also in modern news, literature, and propaganda. Expressions of motifs in text can be most straightforwardly classified as *simple* or *complex*. Simple motif expressions are easy to detect because they almost always appear in a single sentence using the same words as the motif definition itself. However, complex motifs are strongly non-literal and often spread across multiple sentences, thus requiring more context to understand. We propose a baseline system to detect complex motif expressions that have challenged prior work. We used an annotated corpus that identified 992 complex motif expressions of 155 different motifs for training and testing. We tested five different generative approaches that included varying amounts of context: a single sentence baseline (from prior work); a window of 3 or 5 sentences; the entire story; or the entire story with the target sentence identified. We fine-tuned four off-the-shelf open-source LLMs using LoRA under these conditions. Somewhat surprisingly, we report a negative result: our experiments show that in our generative setup more context did not reliably improve the performance of detecting complex motifs, and often hurt fine-tuned models. We speculate on why this might be so and identify directions for future research.

Keywords: folklore, motifs, the Arabian Nights, natural language processing, neural methods, large language models, linguistic annotation

1. Introduction

Motifs are non-commonplace, specific, recurring narrative elements that are often found originally in folk stories and are more generally deployed in culturally inflected materials. Motifs are interesting because they are a compact source of cultural knowledge: many motifs concisely communicate a constellation of related ideas, associations, and assumptions. For example, "troll under a bridge" is a motif common in Western cultures with roots in Scandinavia. To those familiar with the motif, it entails a number of related ideas that are not directly communicated by the surface meaning of the words: the bridge is along the critical path of the hero, and he must cross it to achieve his goal; the troll often lives under the bridge, crawling out to waylay innocent passers-by; the troll charges a toll or demands something for crossing the bridge; the troll is a squatter, not the officially sanctioned master of the bridge; the troll enforces his illegitimate claim through threat of physical violence; and the hero often ends up battling (and defeating) the troll instead of paying the toll.

While motifs usually originate in folkloristic material, they are frequently used in modern discourse, and motif expressions can be easily found in speeches, news reports, press releases, propaganda, books, and movies; indeed, in any type of language where cultural knowledge is deployed. An excellent example of such modern usage is the Islamist motif *Pharaoh*. The Pharaoh appears in stories found in the Hebrew and Christian Bibles and the Qur'an; in those stories, the Pharaoh comes into conflict with Moses and his attempts to free the Hebrews from Egyptian slavery. The Pharaoh is an arrogant and obstinate tyrant who defies the will of God and is punished for it. In modern Islamist extremist narratives, the Pharaoh is a symbol of struggles against anti-Islamic regimes and has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush of the United States, the last of whom Osama bin Laden referred to as the "pharaoh of the century" (Halverson et al., 2011). Further, one must be familiar with Islamic religious and folkloristic traditions to understand the use of this motif in modern language; its meaning is metaphoric and obscure to those not versed in the tradition of the group.

Motif expressions can be most easily classified into two types: simple and complex. Simple motif expressions match the motif definition (found in a motif index) nearly word-for-word. For example, the motif *Mermaid* is found expressed in the Arabian Nights as: *While he was doing this, the sea became disturbed and out from it came mermaids, the sea's daughters, each carrying in her hand a jewel gleaming like a lamp.* (Irwin, 2010, Volume 2, Night 491). On the other hand, the words used in a complex motif expression don't overlap those used in the definition, with the expression often being non-literal and indirect. Complex motif expressions thus presumably need more context and language understanding in order to detect and interpret.

For example, consider the motif *Magic sphere*

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 40–50
11 May 2026.

burns up country. By turning that part of the globe to the sun, one can make any place on earth burn up. This motif is found expressed in the Arabian Nights as:

> *Whoever has the globe can, if he wants, sit inspecting all lands from east to west and whatever part he wants to see, he can do so by turning the globe where he wants and looking into it. He will then have a view of the land and its people as though they were all there in front of him. If he is angry with any city and turns the globe towards the sun with the intention of burning the city to the ground, this is what will happen. As for the kohl case, whoever uses its contents on his eyelids will see all the treasures of the earth.*

Note that, aside from expression complexity, there is a difference in the complexity of the concept of the motif itself: the idea of a *mermaid*—which refers to a single, albeit imaginary, class of things—is in many ways much simpler than the motif of the magic sphere above. Thus motifs themselves can be simple in their conceptual structure, or complex. We will explore this distinction as well in the work.

While simple motif expressions are relatively easy to detect because of their simple form, automatically detecting and interpreting complex motif expressions is challenging, as has been demonstrated in prior work (Alyami and Finlayson, 2026). In particular, we used an annotated corpus developed in that work which identified 992 complex motif expressions of 155 different motifs. Using these data, we explore tackling the task of automatically detecting complex motif expressions in the folkloristic materials by using more context around the target of the non-literal expressions of cultural motifs. We tested five different generative approaches, which used different amounts of context in the prompt: (1) single sentence baseline (from prior work); (2) 3-sentence window of context; (3) 5-sentence window of context; (4) the entire story; and (5) the target sentence plus the entire story. We tested four LLMs, namely: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, google/gemma-3-4b-it, and Qwen/Qwen3-8B.

This paper is structured as follows. We first provide background on motifs and review related computational work (Related Work). We then describe the dataset (Data). We describe the methods we explored, including fine-tuned LLMs using LoRA (Methods). Finally, we discuss the results, the limitations of the work, and possible next steps (Discussion) and we conclude with a list of our contributions (Contributions).

2. Related Work

2.1. Prior Work

Alyami and Finlayson (2026) developed a model to index motifs automatically from *The Arabian Nights*. They built the model using an annotated dataset that consists of 200 motifs extracted manually from *A Motif Index Of The Thousand and One Nights*. The index contains around 5,000 motifs extracted by El-Shamy from the 207 stories of The Arabian Nights (El-Shamy, 2006). Alyami and Finlayson (2026) analyzed the conceptual complexity of the motifs (classifying them as simple or complex), and also analyzed the complexity of the motif expressions (again, simple or complex), resulting in a four-way classification of motif expressions.

First, **Simple Structure** / **Simple Expression** motif expressions are where both the conceptual structure of the motif and the way it is expressed in the narrative are straightforward. These expressions are relatively easy to detect. For example, *Seven daughters* is simply structured and usually expressed using those exact words: *The Almighty provided him with seven daughters. . .* (Irwin, 2010, Volume 3, Night 784).

Second, **Simple Structure** / **Complex Expression** motif expressions are where a conceptually simple motif is expressed in a complex way. In other words, the motif is expressed in the index in only one or a few words, but in the text the motif is expressed indirectly, obscurely, or across many pages. For example, the motif *Resourcefulness* is conceptually simple, but one expression runs *The mamluks delightedly agreed that this was a good plan, and there and then they began to cut logs for the raft and to twist ropes to bind them together. They worked on this for a month, taking back firewood for the princess's kitchen each evening and devoting the rest of the day to the raft.* (Irwin, 2010, Volume 3, Night 766).

Third, **Complex Structure** / **Simple Expression** motif expressions are where the motif is conceptually complex, but still the expression of the motif in the text closely follows what is found in the index, and so keyword retrieval usually suffices to find them. For example, the motif *Apparently dead persons revived when certain thing happens. Proper prince appears, or the like.* is not simply structured but can be expressed almost exactly as is found in the index, as in *You Muslims, you soldiers, have you ever in your lives seen a man die and then come back to life?* (Irwin, 2010, Volume 1, Night 34).

Fourth and finally, **Complex Structure** / **Complex Expression** is the most difficult type of motif expression to find. The motif is conceptually complicated and motif expression is also complex. For example, the motif *What you (deal) to others will*

be done (dealt) back to you is found in the text as follows:

> *I took hold of the horse and mounted it. It didn't move and so I kicked it, and when it still refused to move, I took the whip and struck it. It didn't move and so I kicked it, and when it still refused to move, I took the whip and struck it. As soon as it felt the blow, it neighed with a sound like rumbling thunder and, opening up a pair of wings, it flew off with me, carrying me up into the sky way above the ground. After a time, it set me down on a flat roof and whisked its tail across my face, striking out my right eye and causing it to slide down my cheek. It then left me and I came down from the roof to find the ten one-eyed youths. No welcome to you, they said. Here I am, I replied. I have become like you, and I want you to give me a tray of grime with which to blacken my face and to let me sit with you. No, by God, they said, you may not do that. Get out! (Irwin, 2010, Volume 1, Night 16).*

The motif is found in across approximately 190 words and 10 sentences, where none of the words in the motif definition are found directly in the expression.

Alyami's model was trained on single sentences that were annotated for the presence or absence of particular motifs. The most effective method for detection was a fine-tuned Llama3, achieving an overall F_1 performance of 0.90 for simple motif expressions (across both conceptually simple and complex motifs). Their study shows that this task is still challenging for state-of-the-art LLMs when trying to detect complex motifs, achieving an overall F_1 performance of 0.72.

Other work has also emphasized that automatically detecting and interpreting motif expressions in modern language is a challenging problem (Yarlott and Finlayson, 2016; Yarlott et al., 2022; Yarlott, 2022; Yarlott et al., 2024; Acharya, 2022; Acharya et al., 2024). Prior work defined the task of *motif detection*, which is finding a motif expression in non-folkloristic materials. In that task, even a word-by-word expression of a motif must be further differentiated into MOTIFIC, EPONYMIC, REFERENTIAL, or UNRELATED types (Yarlott et al., 2024). An EPONYMIC usage is the use of the motif as a name; a REFERENTIAL usage is a mention of the motif itself; while a MOTIFIC usage is intended to call to mind the implicit associations of the motif. Here we tackle a different task, which is detecting appearances of a complex motif in the original folkloristic materials, which we will call *detection of complex, non-literal expressions of cultural motifs*. Methods that address this task would allow the automatic mining of positive and negative examples to train and test other stages of motif understanding, such as discovery of novel motifs (i.e., *motif discovery*: Yarlott and Finlayson, 2016), identification of motif usage in non-folkloristic materials (i.e., *motif detection*, as above: Yarlott et al., 2022, 2024), and interpretation of the meaning of motific language (i.e., *motif interpretation*: Acharya, 2022; Acharya et al., 2024).

2.2. Culture in LLMs

Recent research has highlighted the challenges LLMs face in understanding cultural elements from text. Adilazuarda et al. (2024) explained the need for context, or called "thick description" by Geertz (1973). The concept of thick description or context is needed in order to understand a culture as insider. The context needed to analyze the internal or specific small details of that cultural text to understand not only behaviors or certain events, but also cultural, religious, and psychological norms which will make LLMs more culturally aware. On the other hand, "thin description" also coined by Geertz (1973), is not context-aware analysis, but rather when an outsider frames the text without understanding why certain behaviors or habits will be interpreted by cultural insiders differently depending on the situation. Adilazuarda et al. notes that most of the current research on the cultural understanding of LLMs focuses mainly on values and emotions. However, there are many other aspects of cultures text can and should be explored.

The challenge of figurative language understanding across cultures has also been explored more broadly. Kabra et al. (2023) created a multilingual figurative language inference dataset (MABL) across seven culturally diverse non-English languages. Their work showed that figurative expressions are deeply rooted in culturally specific concepts such as food, religion, and events with overlap between languages from the same geographic region. Liu et al. (2024) evaluates Multilingual Large Language Models (mLLMs) to reason using cultural common ground by using proverbs and sayings from many languages as an exploratory method. Proverbs are culturally specific expressions. They collect proverbs and their usage in conversational situations from six languages. They examine distinct mLLMs to assess their capacity to memorize proverbs, use proverbs and sayings to reason in varied situational circumstances, and comprehend proverbs in cross-cultural conversations. They created a dataset called Multicultural Proverbs and Sayings (MAPS) for proverb understanding with conversational context for different languages. Their study found that mLLMs possess knowledge of proverbs and sayings to varying

degrees, but memorizing a proverb does not indicate the ability to reason with it in context. They also found significant culture gaps when reasoning across languages, with performance on English data consistently stronger than other languages, particularly for lower-resource languages such as Bengali and Indonesian. Park et al. (2025) stated that LLMs often succeed in recognizing figurative expressions at the sentence level but their ability to use them coherently in conversation remains uncertain. They showed that even models that recognize figurative expressions at the sentence level fail to use them appropriately in dialogue. Our work extends this line of work to the domain of folkloric cultural motifs where expressions are not only figurative but often span multiple sentences and require deep cultural background knowledge.

With this in mind, our paper evaluates LLMs on a diverse set of motifs that cover different cultural aspects of folktales. The motif index by El-Shamy (2006), which is where we source our motif lists, categorizes motifs based on their cultural, social, and psychological contextual significance. For example, a motif can represent a cultural theme such as *Tabu: eating with left hand.* (El-Shamy, 2006). In Islamic-Arabic culture, it is taboo to eat with your left hand due to the cultural habits. In the narrative, the motif is expressed in Irwin (2010) as follows:

> *I went and prepared the necessary food, drink, and so on, which I then presented to him, inviting him to eat in the Name of God. He went to the table and stretched out his left hand, after which he ate with me. This surprised me, and when I had finished, I washed his hand and gave him something to dry it with. I then sat down to talk, after I had offered him some sweetmeats. 'Sir,' I said to him, 'you would relieve me of a worry were you to tell me why you ate with your left hand. Is there perhaps something in your other hand that causes you pain?*

Motifs may also have social significance: *No low rank person would be sitting down while addressing high rank* is an example of how motifs can represent the social theme in the narrative (El-Shamy, 2006). It is expressed in Night 620 in Irwin (2010) as follows:

> *Uthman was both stupid and conceited, and when he arrived at Judar's palace he saw a eunuch seated on a chair in front of the door. This man did not get up on his arrival, and in spite of the fact that there were fifty men with 'Uthman, it was as though no one had come. 'Uthman went up to him and said: 'Slave, where is your master?' 'In the house,' the eunuch replied, and as he spoke he continued to lounge on his chair.*

Third, a *Clothes make the man* is an example of how motifs can represent psychological themes, as follows:

> *The weaver went along and saw magnificently dressed people receiving fine foods and being treated with respect by the host because of their splendid clothes. He said to himself: 'Were I to change my trade for one that would be of less trouble, more prestigious and more rewarding, I could collect a lot of money and buy clothes like these. I would then become important; people would respect me and I would be like these other's.*

A related concern is that LLMs are biased to certain cultures, especially when prompting in English. The reason behind that is that the datasets used to train these LLMs are mostly in English and represent Western cultures and values (Tao et al., 2024; Johnson et al., 2022; Atari et al., 2023). The work we present here is an example of moving beyond the Western tradition, in that the motif index used in our data is specific to the Middle East, and contains mainly Muslim cultural material.

3. Data

We began with the data that was collected and used in prior work, by Alyami and Finlayson (2026). In that work, Alyami and Finlayson used the *Motif Index of The Thousand and One Nights* (El-Shamy, 2006), which contains around 5,000 motifs extracted by El-Shamy from the 207 stories of The Arabian Nights. These motifs overlap with motifs found in the Thompson Motif Index (TMI), while adding new motifs specific to the Arabic content of the stories (Thompson, 1955-1958). El-Shamy generally followed the TMI classification scheme (i.e., 23 themes, each identified by a letter), but he also added categories and information corresponding to dimensions of cultural, social, and psychological contextual significance. The annotated dataset comprised 58,450 annotated sentence-motif pairs, of which 2,670 were positive examples. These annotations used 200 unique motifs. There were examples of motif expressions from all four types (simple-simple, simple-complex, etc.). From these data, we selected only the complex expressions. This subset of the data comprised 992 sentences with positive examples. It represents 155 unique motifs. Note that Table 1 reports the number of unique motifs per split (112 for training, 24 for validation, and 19 for testing) and not only the number of motif expressions. The full set of 992 complex

Conceptual Complexity ↓	Uniq. Motifs	Motif Exps.	Training / Validation / Testing
Simple	108	689	81 / 17 / 10
Complex	47	303	31 / 7 / 9
Overall	155	992	112 / 24 / 19

Table 1: Number of unique motifs, number of complex motifs expressions (positive examples), and the training / validation / testing set sizes (in number of unique motifs) broken down by the conceptual complexity of the motifs.

Data	Count
Unique Stories	64
Total Sentences in Unique Stories	1,393
Total Tokens in Unique Stories	34,513
+Examples in the Index	266
Motifs w/ 1 +Example in the Index	175
Motifs w/ >1 +Example in the Index	26
Sentences-Motif pairs, +Examples	2,670
+Examples Total Tokens	88,523

Table 2: Data

motif expressions is distributed across these splits in a balanced way. We paired these with an equal number of random sentences that didn't contain any motif expression, resulting in 1,984 total examples[1]. The dataset is split into training, validation, and test sets. Table 1 shows the breakdown of motifs and motif expressions by the conceptual complexity of the motifs (simple or complex). Each training, testing, or validation comprises a unique list of motifs so that each model is tested on an unseen list of motifs. Additional details on the data are found in Table 2.

Annotation of complex motif expressions is inherently challenging, as it can be a tedious and, at times, challenging task depending on the complexity of the data. The task is to label sentences as True Positive (TP) when the motif is present and False Positive (FP) when it is not, and the sentences and motifs fall into four categories: Simple Expression, Complex Expression, Simple Structure, and Complex Structure. It is essential to ensure every element of the motif is present when annotating, as annotators might feel tempted to label something as TP because they want it to be, rather than because it truly meets the criteria — a bias that often occurs after a long string of FPs, and mislabeled data can pollute the dataset, leading to inaccurate model training. These sentences are isolated from their broader context, which can cause difficulties. For example, proper names may appear without their associated titles, leading to potential mislabeling. The difficulty of the annotation task itself provides important context for interpreting model performance if human annotators also struggle with complex motif expressions, model scores below human-level performance are expected rather than surprising.

[1]The code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/VKJEGZ. The original full dataset is available from the authors Alyami and Finlayson (2026).

4. Methods

We evaluated five different generative approaches to solving the task. In all five approaches, the task is formulated as a binary classification per text-motif pair: given a specific motif and a text. The model must predict whether that particular motif is expressed in the text. Each text is evaluated against exactly one motif at a time. The baseline is a single-sentence approach (§4.1) as demonstrated in Alyami and Finlayson (2026). Then we experimented with a three sentence context window (target sentence plus one sentence before and after), a five sentence context window, the entire story, and finally, the target sentence coupled with the entire story.

We experimented with four open-source LLMs fine-tuning using Low-Rank Adaptation (LoRA): Mistral (Mistral-7B-Instruct-v0.3) (May 2024), Llama (Llama-3.1-8B-Instruct) (July 2024) and Google (gemma-3-4b-it) (March 2025), and Qwen (Qwen3-8B) (May 2025) (Mistral AI, 2024; Meta, 2024; Gemma Team, et al., 2025; Qwen Team, et al., 2025). We used the training/validation/test dataset we showed in Table 1.

We fine-tuned all LLMs using Low-Rank Adaptation (LoRA), a parameter-efficient method (Hu et al., 2022). The goal was to see an improvement in classifying whether a text contains a motif or not with different amounts of context (3-sentence window, 5-sentence window, entire story, or target sentence plus entire story). We fine-tuned the LLMs using the same dataset we used in fine-tuning embedding models. Each training example consists of motif-sentence pairs as input, target output (Yes or No) only. We set epochs to 5, the learning rate to 10^{-4}, batch size to 16, temperature to 0, and new tokens to 1 only in order to force the models to be deterministic.

Illustrative Example

To show how each of the five context methods is applied. We walk through an example using the motif *Eblis: born as one of the fourteen children of*

Khalit and Malit. He disobeyed his father by refusing to marry one of his seven twin-sisters, and was transformed into a worm (which became Eblis). The motif is expressed in Night 493 in Irwin (2010) as follows:*Then, when it was the four hundred and ninety-third night, SHE CONTINUED:I have heard, O fortunate king, that when Buluqiya had told King Sakhr the full story of his wanderings from beginning to end, the king was filled with astonishment and ordered his servants to fetch tables, which they spread with cloths. Then they brought plates of red gold, of silver and of copper. On some of them were fifty cooked camels and on others twenty, while some contained fifty sheep. In all there were one thousand, five hundred plates, and when Buluqiya saw that he was amazed.The company then ate, as did Buluqiya, who, when he had had enough, gave thanks to Almighty God. After that, the food was removed and was replaced with fruit. When they had all finished eating, they called down praises on Almighty God and blessings on His Prophet, Muhammad, may God bless him and give him peace. Buluqiya was surprised to hear the name of Muhammad and he asked the king if he might put some questions to him. 'Ask what you want,' the king told him, and so he said: 'O king, what are you? What is your origin and how do you come to know of Muhammad, so that you call down blessings on him and love him?''Buluqiya,' replied the king, 'Almighty God created hellfire in seven layers, one on top of the other, each separated by the distance of a thousand years' journey. The first of these layers is called Jahannam and it has been prepared for those Muslims who disobey God's commands and die without having repented. The second layer is called Lazan and this is prepared for the unbelievers, while the third is Jahim, prepared for Gog and Magog. The name of the fourth layer is al-Sa'ir and this is for the people of Iblis; the fifth is called Saqar and is for those who abandon prayer. The sixth is al-Hutama and is for Jews and Christians, while the seventh is al-Hawiya, which has been prepared for the hypocrites. These are the seven layers.''I suppose,' said Buluqiya, 'that the punishments of Jahannam are easier to bear than all the others as it is the uppermost layer.' The king agreed with this, but added: 'In spite of that, Jahannam contains a thousand mountains of fire, in each of which there are seventy thousand valleys, each containing seventy thousand cities of fire.In each of these cities there are seventy thousand fiery castles, with seventy thousand fiery rooms in each, and each room contains seventy thousand couches of fire, with seventy thousand forms of torment in every one of them. None of the other layers, however, have any lighter punishments than these, as this is the first layer, while as for the other layers, only God Almighty knows the number of their torments.'When Buluqiya heard what the king had to say he collapsed in a faint, and when he recovered he burst into tears and said: 'O king, how then will it be with us?' 'Have no fear,' said the king, 'for you must know that the fire will not burn anyone who loves Muhammad, and for his sake, may God bless him and give him peace, such a man will be freed, while hellfire will flee from all who follow his religion. As for us, Almighty God created us from fire, and the first beings that He created in Jahannam were two of his host, the first called Khalit and the second Malit. Khalit was shaped like a lion and Malit like a wolf. Malit's tail was feminine, piebald in colour, while Khalit's was masculine, in the shape of a tortoise, and was a twenty-year journey in length. God then ordered these two tails to join together and copulate, and from them were born snakes and scorpions who live in hellfire and, having reproduced and multiplied, are used by God to torture those who enter it. God then ordered the two tails to copulate a second time, and when they did this, Malit's tail was impregnated by the tail of Khalit and gave birth to seven males and seven females. These were nurtured until they grew up, and when they had done so, the females were married to the males. All but one of them were obedient to their father; the one who disobeyed became a worm and this worm is Iblis, may God Almighty curse him. He had been one of the cherubim, serving God until he was raised to heaven, where he found favour with Him and became the leader of the cherubim.'Morning now dawned and Shahrazad broke off from what she had been allowed to say.* The story contains the following sentences (numbered for reference):

> s_{-2}: Malit's tail was feminine, piebald in colour, while Khalit's was masculine, in the shape of a tortoise, and was a twenty-year journey in length.
> s_{-1}: God then ordered these two tails to join together and copulate, and from them were born snakes and scorpions who live in hellfire and, having reproduced and multiplied, are used by God to torture those who enter it.
> s_0: **God then ordered the two tails to copulate a second time, and when they did this, Malit's tail was impregnated by the tail of Khalit and gave birth to seven males and seven females.**
> s_{+1}: These were nurtured until they grew up, and when they had done so, the females were married to the males.
> s_{+2}: All but one of them were obedient to their father; the one who disobeyed became a worm and this worm is Iblis, may God Almighty curse him.

Single Sentence (§4.1): Only s_0 is provided to the model without any surrounding context.
3-Sentence Window (§4.2): The model receives s_{-1}, s_0, and s_{+1}. This provides minimal local context.
5-Sentence Window (§4.2): The model receives s_{-2}, s_{-1}, s_0, s_{+1}, and s_{+2}. This provides a broader local context.
Entire Story (§4.3): The full story text is provided. In this case, the model must determine whether the motif appears anywhere in the story.
Target Sentence + Entire Story (§4.4): The full text of the story is provided as background context. We ask the model to classify the target sentence s_0 using the entire story as background.

4.1. Single Sentence

In this baseline, we fine-tuned the four LLMs on the training and validation sets shown in Table 1, with the goal of teaching the models to understand the smallest possible descriptions of motifs. We fed the motif a single positive or negative sentence and the models were asked to classify the sentence. The requested answer is *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the sentence.
Rules: Answer ONLY "Yes" or
"No". Do not explain.
Motif: <Motif>
Sentence: <Sentence>
Answer:
```

4.2. Window of Target Text (3- or 5-sentence)

Here, we add either two or four sentences of context (one or two sentences before and after), and asked the models to classify if the target sentence with this window contains the motif or not. The answer is again only *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the target text.
Context is provided ONLY to
resolve ambiguity Rules: Answer
ONLY "Yes" or "No". Do not
explain.
Motif: <Motif>
Context Before: <Pre Target
sentence>
Target Sentence: <Target
Sentence>
Context After: <Post Target
Sentence>
Answer:
```

4.3. Entire Story

For maximum context we feed the entire story in with the motif. Following previous modes where the model is asked to classify if the story contains the motif or not by only answering *Yes* or *No* using the following prompt:

```
Task: Decide if the motif is
present anywhere in the story
text.
Rules: Answer ONLY "Yes" or
"No". Do not explain.
Motif: <Motif>
Story Text: <Story>
Answer:
```

4.4. Target Sentence plus Entire Story

In this approach, we fed the models the sentence to be classified, plus the entire story for complete context, and asked the models to classify whether the story contains the motif in the target sentence or not. Again, the requested answer is *Yes* or *No*. We used the following prompt:

```
Task: Decide if the motif is
present in the target sentence.
The story is background only.
Hard rule:
- Answer 'Yes' ONLY if the TARGET
sentence itself clearly expresses
the motif.
- If the motif is only elsewhere
in the story, or only implied
weakly, answer 'No'.
- When in doubt, answer 'No'.
Answer ONLY 'Yes' or 'No'. Do
not explain.
Motif: <Motif>
Story (background): <Story>
Target Sentence: <Target
Sentence>
Answer:
```

5. Discussion

The results are shown in Table 3. In our setup, increasing context does not consistently improve performance. Fine-tuned models often degrade as context length increases, especially in the entire-story mode. However, some zero-shot models benefit from added context. We suspect that this is because motifs are a relatively small part of the text, and the models have trouble focusing on such a small portion. Compared to the entire story, the LLMs, in most cases, failed to detect the complex motif expressions for both simple and complex motif conceptual complexity classes. While all models failed in most cases, Mistral-FT shows significantly weaker performance when increasing the context. This might suggest a direction when comparing

Conceptual Complexity ↓		Complex Expressions				
		Single Sentence	±1 Window	±2 Window	Entire Story	Target Sentence+ Entire Story
Simple	Mistral-Zero-shot	0.31 (0.66 / 0.20)	0.44 (0.73 / 0.31)	0.39 (0.68 / 0.27)	**0.65** (0.55 / 0.80)	0.45 (0.69 / 0.34)
	Mistral-FT	0.65 (0.74 / 0.58)	0.52 (0.67 / 0.43)	0.50 (0.71 / 0.39)	0.10 (0.33 / 0.06)	0.26 (0.56 / 0.17)
	LLama3-Zero-shot	0.64 (0.57 / 0.73)	**0.65** (0.60 / 0.70)	0.59 (0.59 / 0.59)	0.63 (0.59 / 0.67)	0.44 (0.55 / 0.37)
	LLama3-FT	**0.73** (0.81 / 0.67)	0.50 (0.73 / 0.39)	0.59 (0.82 / 0.46)	0.60 (0.63 / 0.57)	0.46 (0.44 / 0.47)
	Gemma3-Zero-shot	0.65 (0.53 / 0.83)	0.36 (0.50 / 0.29)	0.22 (0.32 / 0.17)	0.63 (0.49 / 0.86)	0.60 (0.48 / 0.81)
	Gemma3-FT	0.41 (0.59 / 0.31)	0.34 (0.51 / 0.26)	0.39 (0.61 / 0.29)	0.62 (0.53 / 0.74)	0.63 (0.55 / 0.73)
	Qwen-Zero-shot	0.45 (0.67 / 0.34)	0.62 (0.71 / 0.56)	**0.64** (0.65 / 0.63)	0.55 (0.72 / 0.44)	**0.64** (0.56 / 0.75)
	Qwen-FT	0.58 (0.72 / 0.49)	0.29 (0.65 / 0.19)	0.21 (0.53 / 0.13)	0.62 (0.70 / 0.56)	0.57 (0.49 / 0.66)
Complex	Mistral-Zero-shot	0.46 (0.91 / 0.31)	0.49 (0.85 / 0.34)	0.37 (0.73 / 0.25)	0.53 (0.67 / 0.44)	0.30 (0.100 / 0.17)
	Mistral-FT	0.72 (0.91 / 0.60)	0.68 (0.86 / 0.56)	0.72 (0.90 / 0.59)	0.06 (0.50 / 0.03)	0.14 (0.33 / 0.09)
	LLama3-Zero-shot	0.64 (0.85 / 0.51)	0.64 (0.89 / 0.50)	0.61 (0.88 / 0.47)	0.47 (0.63 / 0.38)	0.21 (0.60 / 0.13)
	LLama3-FT	0.70 (0.86 / 0.59)	0.60 (0.93 / 0.44)	0.52 (0.86 / 0.38)	0.53 (0.57 / 0.50)	0.38 (0.38 / 0.39)
	Gemma3-Zero-shot	0.57 (0.71 / 0.47)	0.36 (0.67 / 0.25)	0.36 (0.62 / 0.25)	0.62 (0.49 / 0.84)	0.57 (0.45 / 0.78)
	Gemma3-FT	0.59 (0.73 / 0.50)	0.43 (0.71 / 0.31)	0.41 (0.44 / 0.38)	0.59 (0.48 / 0.75)	0.51 (0.44 / 0.61)
	Qwen-Zero-shot	0.74 (0.91 / 0.63)	**0.81** (0.92 / 0.72)	**0.76** (0.91 / 0.66)	**0.65** (0.94 / 0.50)	**0.59** (0.67 / 0.52)
	Qwen-FT	**0.84** (0.80 / 0.88)	0.75 (0.88 / 0.66)	0.75 (0.88 / 0.66)	0.60 (0.61 / 0.59)	0.43 (0.47 / 0.39)
Overall	Mistral-Zero-shot	0.36 (0.75 / 0.23)	0.46 (0.77 / 0.32)	0.38 (0.69 / 0.26)	**0.63** (0.57 / 0.69)	0.42 (0.73 / 0.29)
	Mistral-FT	0.67 (0.78 / 0.58)	0.57 (0.73 / 0.47)	0.57 (0.78 / 0.45)	0.09 (0.36 / 0.05)	0.23 (0.50 / 0.15)
	LLama3-Zero-shot	0.64 (0.62 / 0.66)	0.65 (0.66 / 0.64)	0.60 (0.65 / 0.55)	0.59 (0.60 / 0.58)	0.39 (0.56 / 0.30)
	LLama3-FT	**0.72** (0.82 / 0.64)	0.53 (0.79 / 0.40)	0.57 (0.83 / 0.43)	0.58 (0.62 / 0.55)	0.44 (0.42 / 0.45)
	Gemma3-Zero-shot	0.63 (0.56 / 0.72)	0.36 (0.54 / 0.27)	0.26 (0.40 / 0.20)	0.62 (0.49 / 0.85)	0.59 (0.47 / 0.80)
	Gemma3-FT	0.47 (0.64 / 0.37)	0.37 (0.57 / 0.27)	0.40 (0.53 / 0.31)	0.61 (0.51 / 0.75)	0.59 (0.52 / 0.70)
	Qwen-Zero-shot	0.55 (0.76 / 0.43)	**0.68** (0.78 / 0.61)	**0.67** (0.71 / 0.64)	0.58 (0.78 / 0.46)	**0.63** (0.58 / 0.68)
	Qwen-FT	0.68 (0.76 / 0.61)	0.47 (0.77 / 0.33)	0.42 (0.73 / 0.29)	0.61 (0.67 / 0.57)	0.53 (0.49 / 0.59)

Table 3: Overall System Evaluations: F_1 (precision / recall).

two texts of different lengths. One approach could be to summarize the target text. However, it will be challenging for a system to summarize a story while ensuring that the relevant motif remains in the summary; to ensure this in all cases, the system will need to already have the ability to detect the motif. Another possible approach is to increase the length of the motif with more description, so that the model has enough indications of where the motif could be located. In general, our observation that more context actually hurts detection for a small piece of target text (at least under our experimental setup) is aligned with a prior work that discusses the 'Needle in a Haystack' tasks, where they show that smaller relevant contexts degrade LLM performance (Bianchi et al., 2025).

When looking at the precision and recall breakdown, we notice that the performance drop with increasing context is driven by a collapse in recall rather than precision. On the other hand, in the fine-tuned models, precision remains relatively stable across context windows. For instance, Mistral-FT maintains precision around 0.67–0.91 across single-sentence and windowed modes but the recall drops sharply when moving to entire-story mode (e.g., from 0.58 at single sentence to just 0.06 for simple-conceptual motifs, and from 0.60 to 0.03 for complex-conceptual motifs). This implies that fine-tuned models become too cautious with additional information. This effect is consistent across both simple and complex conceptual motifs but particularly noticeable for complex-concept motifs under the entire-story mode. Interestingly, the windowed modes (±1 and ±2) preserve much of the single-sentence performance. For instance, Mistral-FT holds F_1=0.68–0.72 across windows for complex-concept motifs. This suggests that a small amount of local context does not hurt but a large context window is harmful. By contrast, zero-shot models show the opposite pattern where the recall rises substantially at the entire-story level (e.g., Mistral-Zero-shot recall reaches 0.80 for simple-concept motifs and 0.44 for complex-concept motifs on the entire story compared to just 0.20 and 0.31 at single sentence mode). On the other hand, precision suffers. This implies that when models are fine-tuned on short-span text they learn to look for specific words in short text. However, when a motif is spread across a long story, they suffer when trying to detect these complex motif expressions.

Comparing fine-tuned and zero-shot models shows that LoRA fine-tuning with a small data size gives mixed results. For short contexts, fine-tuning helps in some cases. For instance, LLama3-FT achieved F_1=0.73 which outperforms LLama3-Zero-shot at F_1=0.64 for simple-concept motifs, and Qwen-FT achieved F_1=0.84, outperforming Qwen-Zero-shot at F_1=0.74 for complex-concept motifs at the single-sentence level. However, this is not consistent across all models. Gemma3-FT, for instance, underperforms Gemma3-Zero-shot

in nearly every mode. On the other hand, for longer contexts, fine-tuning is clearly harmful. Zero-shot models consistently outperform fine-tuned models, especially on the entire-story mode. For example, Mistral-Zero-shot reaches F_1=0.65 for simple-concept motifs versus just 0.10 for Mistral-FT, and Qwen-Zero-shot leads complex-concept motifs with F_1=0.65 versus 0.60 for Qwen-FT. Overall, LLama3-FT achieves the best single-sentence F_1 of 0.72, but Qwen-Zero-shot is the strongest model across all wider-context modes (F_1=0.68, 0.67, and 0.63 for ±1 window, ±2 window, and Target Sentence + Entire Story, respectively). Mistral-Zero-shot is leading on the entire-story mode at F_1=0.63. This may suggest that fine-tuning on limited data causes models to focus on short-span patterns, which could reduce their ability to reason over longer contexts. Collecting more training data and exploring other fine-tuning strategies could help address this gap.

6. Contributions

We fine-tuned four open-source LLMs using LoRA using 992 complex motif expressions using varying amounts of context: a single sentence baseline (from prior work); a window of 3 or 5 sentences; the entire story; or the entire story with the target sentence identified. The most effective model was Llama3, achieving an overall F_1 performance of 0.72 in a single sentence context. The study shows that this task remains challenging for state-of-the-art LLMs, and that additional context does not provide a consistent benefit across models and prompting modes (e.g., sentence, windows, or the entire story), with Qwen being the most effective model in the wider-context settings, achieving an overall F_1 score of 0.63. We release code and data for reproducing this study [2].

7. Limitations

The first limitation is that while our experiments show that increasing context does not improve performance under these specific experimental conditions, it is hard to generalize from this to *all* potential experimental conditions. Therefore, although our results are relatively straightforward and suggest that more context does not help these kinds of models, we cannot completely eliminate the possibility that some experimental setup using generative models of this kind won't respond positively to additional context.

A second limitation of this work is the amount of annotated data available to train the system. For future work, we believe that collecting more complex expression motifs would be valuable, since until now there is no source for these motif expressions except the prior work of (Alyami and Finlayson, 2026). While their data contains a small number of annotated examples (155 motifs with 992 positive examples), we believe that continuing to expand the annotated data will allow the development of even more capable models. It would also be useful to have a way of precisely and ideally automatically assessing the complexity of both motif conceptual structure and motif expressions. Right now, complexity judgments are done manually, and a more precise standard would allow more careful separation of hard and easy examples for training and testing the systems. Another limitation is that we did not evaluate the model when there is another motif in the same text. In the current methodology, each text is evaluated against only one motif. This is an interesting open problem to see how state-of-the-art models perform when a text contains motif X and also contains a related but distinct motif Y. Not only that, but also how these modes lead to clear boundaries between these related motifs. We leave this robustness analysis to future work.

[2]The code and data may be downloaded from https://doi.org/10.34703/gzx1-9v95/VKJEGZ.

8. Acknowledgments

This work was supported in part by a Saudi Arabian Cultural Mission Fellowship to Ibrahim H. Alyami from the College of Computer Science and Information Systems at Najran University, Saudi Arabia [grant number 443-16-40].

9. References

Anurag Acharya. 2022. *Integrating Cultural Knowledge into Artificially Intelligent Systems: Human Experiments and Computational Implementations*. Ph.d. dissertation, Florida International University.

Anurag Acharya, Diego Estrada, Shreeja Dahal, W Victor H Yarlott, Diana Gomez, and Mark Finlayson. 2024. Discovering implicit meanings of cultural motifs from text. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)*, pages 46–56.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.

Ibrahim H. Alyami and Mark A. Finlayson. 2026. Automated motif indexing on the arabian nights. *arXiv preprint:2603.19283*.

Mohammad Atari, Mona Xue, Peter Park, Damián Blasi, and Joseph Henrich. 2023. Which humans? *PsyArXiv*.

Owen Bianchi, Mathew J Koretsky, Maya Willey, Chelsea X Alvarado, Tanay Nayak, Adi Asija, Nicole Kuznetsov, Mike A Nalls, Faraz Faghri, and Daniel Khashabi. 2025. Hidden in the haystack: Smaller needles are more difficult for llms to find. *arXiv preprint:2505.18148*.

Hasan M. El-Shamy. 2006. *A Motif Index of The Thousand and One Nights*. Indiana University Press, Bloomington and Indianapolis.

Clifford Geertz. 1973. *The interpretation of cultures*. New York: Basic Books.

Gemma Team, et al. 2025. Gemma 3 technical report.

Jeffry Halverson, Steven Corman, and H Lloyd Goodall. 2011. *Master narratives of Islamist extremism*. Palgrave Macmillan, New York.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Robert Irwin. 2010. *The Arabian Nights: Tales of 1,001 Nights*, volume 1-3. Penguin, London, UK.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico.

Meta. 2024. Llama-3.1-8b-instruct. Hugging Face model repository.

Mistral AI. 2024. Mistral-7b-instruct-v0.3. Hugging Face model repository.

Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. Fluid qa: A multilingual benchmark for figurative language usage in dialogue across english, chinese, and korean. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30268–30282, Suzhou, China.

Qwen Team, et al. 2025. Qwen3 technical report.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Stith Thompson. 1955-1958. *Motif-Index of Folk-Literature, Volumes 1-6: A Classification of Narrative Elements in Folk Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*. Indiana University Press.

Victor Yarlott. 2022. *Communicating with Culture: How Humans and Machines Detect Narrative Elements*. Ph.d. dissertation, Florida International University.

W Victor Yarlott, Anurag Acharya, Diego Castro Estrada, Diana Gomez, and Mark Finlayson. 2024. Golem: Gold standard for learning and evaluation of motifs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7801–7813.

W Victor H Yarlott and Mark A Finlayson. 2016. Learning a better motif index: Toward automated motif extraction. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

W Victor H Yarlott, Armando Ochoa, Anurag Acharya, Laurel Bobrow, Diego Castro Estrada, Diana Gomez, Joan Zheng, David McDonald, Chris Miller, and Mark A Finlayson. 2022. Finding trolls under bridges: Preliminary work on a motif detector. *arXiv preprint:2204.06085*.

Artful Writing, Authentic Emotions: Distinguishing Human-Written from LLM-Generated Metaphors by Annotation and Classification

Michaela Regneri,[1] **Nooshin Aghajari,**[2] **Thomas Kroedel**[2]
[1] Department of Informatics, [2] Department of Philosophy
University of Hamburg, Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

We analyze differences between human-written and automatically generated metaphors. Using two syntactically standardized datasets containing novel metaphors from poetry and science communication, we generate new figurative expressions with LLMs that describe the same concepts as human-written texts. Using crowdsourcing, we conduct extensive annotation across multiple dimensions (e.g., writing quality and creativity) and ask annotators to judge whether the metaphor was generated automatically. For the poetry set, we also asked annotators for the emotions conveyed by the metaphor. We find that, consistent with prior work, the authorship of scientific metaphors is difficult to determine. However, our results reveal that human-written poetic metaphors stand out by their capacity to convey emotion. We also analyze which types of metaphors are merely *perceived as human*. Finally, we show that, while human annotators cannot distinguish human from machine metaphors, automated approaches achieve high accuracy in identifying human writers, which suggests substantial differences in text structure.

1. Introduction

Metaphors are pervasive in human language. While conventional metaphors (e.g., *Time is money*) have become deeply intertwined with language use and are also reflected in dictionary entries (Lakoff and Johnson, 2003), novel (i.e., non-conventional) metaphors are a highly productive part of language. The inventive use of non-literal figures of speech is remarkable in multiple respects: from a cognitive and communicative perspective, new figurative language images are both an authorial creative effort and a communicative challenge for the reader. From a research perspective, it is thus interesting that figurative language can effectively convey emotions (as in poetry) or aid understanding of complex, abstract concepts (as in science communication).

With the advent of Large Language Models (LLMs), questions surrounding the successful use of figurative language become even more complex. While the ability to use metaphors appears to require empathy, for which LLMs lack the capacity, research has shown that LLMs can produce figurative language that humans cannot tell apart from real poetry (Porter and Machery, 2024; Wang et al., 2025).

Our study analyzes the differences between human-written and automatically generated metaphors in greater detail. We first collect two datasets of human-generated novel metaphors. One dataset is derived from poetry and allows to analyze the metaphor's inherent purpose in conveying emotion. The second dataset comprises metaphors from science communication that explain complex concepts through concrete, tangible imagery. We standardize both datasets to the form of analogies ("X is like a Y") to ensure that grammatical variation does not confound our semantic analysis. In a second step, we generate metaphors by using different LLMs. These metaphors describe the same concepts (the X in the analogy) as the human-written metaphors do. We prompt LLMs either to produce emotion-conveying poetic analogies or explanatory metaphors for science communication. To identify features that distinguish human-written and machine-generated metaphors, we annotate all 200 analogies via crowdsourcing and then assess the correlation between the annotations and the metaphors' origins. In a last step, we train automated classifiers on both the metaphors themselves and their annotations, and evaluate how easily the metaphor origins can be identified.

Our main contributions are as follows:

- We provide an extensively annotated

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 51–76
11 May 2026.

dataset of novel metaphors with poetic and scientific metaphors. The metaphors are standardized in analogy syntax, and they comprise human-written and machine-generated examples, matched to each other by a corresponding concept.

- We show that human-written metaphors are distinguished by their less polished writing and, especially, their ability to *convey emotions*: human-written poetry conveys emotions more effectively than automatically generated ones.
- We also analyze what makes metaphors *appear human*. One of the relevant features directly relates to emotion conveyance, namely the annotators' ability to assess the analogy's underlying emotion.
- Regarding the distinction of human-written and LLM-generated metaphors, we show that this is basically impossible for human annotators, possible by automated classification on poetic metaphors, and very accurate using LLMs.

Our results contribute to understanding the subtle differences between the surprisingly large capacities of LLMs for using figurative language and their limits in using metaphors for emotion-driven communication. Our annotated dataset is publicly available.[1]

2. Background and Related Work

Research on metaphor theory and computational methods for understanding it predates LLMs. With this new technology available, new foundations for metaphor identification, creation, and interpretation emerge, raising broader questions about how to identify language generated automatically. Recent work evaluates state-of-the-art models' ability to distinguish metaphorical expressions from literal language and semantic anomalies (Neidlein et al., 2020), and examines how modeling strategies, such as prompt engineering (Kramer, 2025; Jia and Li, 2024), fine-tuning (Haagsma and Bjerva, 2016), retrieval-augmented generation (Fuoli et al., 2025), and detection frameworks (Wang et al., 2025; Lin et al., 2024) improve performance in metaphor identification (Choi et al., 2021; Jia et al., 2025). Despite good results on detection and generation benchmarks, growing evidence suggests that LLMs often rely on superficial cues rather than genuine understanding of the analogy (Sanchez-Bayona and Agerri, 2025), raising new questions about LLMs' capacities to reason about metaphors. While we also use LLMs for classification (among other algorithms), our approach focuses on identifying human-written text rather than identifying figurative language as such.

Beyond identification, recent work explores LLMs' capacity for metaphor generation. Models can create domain-specific metaphors (Stowe et al., 2021; Shou et al., 2024) and metaphors for science communication (Kim et al., 2023). Relating to these approaches, we will analyze whether the generated metaphors share communicative characteristics with human-written ones, and how scientific metaphors differ from poetic ones.

There are multiple datasets containing figurative language, e.g., the LCC Metaphor dataset (Mohler et al., 2016) with metaphors in multiple languages, along with scores for novelty, metaphoricity, affect, and target and source domains. While most other corpora are designed to distinguish literal from figurative expressions, several also annotate the conceptual mappings expressed by the metaphors (e.g., Dodge et al., 2015; Gordon et al., 2015; Shutova and Teufel, 2010). Like in our work, crowdsourcing is often used to collect metaphor annotations, e.g., for metaphoricity (Hovy et al., 2013; Jang et al., 2015; Pedinotti et al., 2021), aptness (Bizzoni and Lappin, 2019), novelty (Parde and Nielsen, 2018; Lugli and Strapparava, 2024; Do Dinh et al., 2018), or to elicit the metaphors (Zayed et al., 2019).

Our work is also located within the broader context of manually or automatically identifying LLM-generated language. The results are mixed: Automatically generated medical student essays can be reliably identified even by medical laypeople (Doru et al., 2025), and LLMs are perceived as less creative than very creative humans (Bellemare-Pepin et al., 2026). In other studies, LLMs do not differ from

[1] https://osf.io/3bgn4/overview?view_only=addb6107053149e387df54647a30f063

average humans in story-telling creativity (Orwig et al., 2024), and LLM-generated poetry is not distinguishable from human-authored poems (Porter and Machery, 2024; Wang et al., 2025). Metaphor interpretations by LLMs are even rated as more useful than those by humans (Ichien et al., 2024). Automated approaches for identifying LLMs as authors are generally more reliable, probably due to structural syntactic differences between LLM-generated and human-written text (Zamaraeva et al., 2025), but they can also be deceived (Shahriar et al., 2025). We extend prior work by analyzing metaphors and the criteria establishing human authorship, using analogy syntax to avoid purely grammatical distinctions.

To the best of our knowledge, we present the first study to investigate which features distinguish human-written metaphors in both the poetry and science domains, both in annotation and classification, and to analyze why metaphors are *perceived* as human-produced. In particular, we highlight the importance of conveying emotion in poetic language.

3. Dataset

Our dataset of novel analogies has two parts: A **poetry** dataset and a **science** communication dataset, both containing human-written and LLM-generated figurative language.

The datasets contain metaphors in the form of analogies. The analogy syntax helps us to keep the expressions comparable and avoid confounding factors (e.g., differences in grammatical structure). Analogies are comparisons of the form "X *(the target)* is like a Y *(the source)*", possibly followed by an explanation that elaborates on the properties of the source that are transferred to the target. While conceptual metaphor theory (Lakoff and Johnson, 2003) emphasizes that metaphors are more deeply intertwined with language than analogies, structural mapping theory supports the view that, particularly for novel metaphors (which often require explanation), they can be understood as analogies. (Bowdle and Gentner, 2005). Just like metaphors, analogies can vary in their degree of novelty: "Life is like a rollercoaster" uses "rollercoaster" as a source, which has a dictionary entry for its non-literal meaning. On the other hand, "Love is like a wanderer who enters the house without knocking" (from our data) is a novel analogy. In this paper, we will use the terms "metaphor" and "analogy" interchangeably.

For the poetry dataset, we adopt the approach of Ichien et al. (2024), who translated Serbian poetry into English to elicit novel metaphors. We extract metaphorical comparisons from Persian and German poetry, as well as 3 originally English metaphors. We rephrase each metaphor as an "X is like a Y" analogy if it is not already phrased that way. E.g., if the text states "My life is an empty notebook", we would record it in our dataset as "My life is *like* an empty notebook", recording "my life" as the target. If the poem contained an explanation, we include it in the analogy. Overall, we collected 20 examples. Two expert annotators noted the primary emotions conveyed by the metaphors, drawing from 6 types: *love, pain, longing, fear, pleasure, regret*, with each analogy assigned one or two emotion labels.

For the science dataset, we collect 20 analogies that explain scientific concepts across disciplines, using the Deep Research component of various LLMs, and manually verify their origin. We include only metaphors whose figurative meanings are not defined in dictionaries. All texts are originally written in English; therefore, the metaphors differ from the poetry set not only in purpose and sentiment but also in the degree of novelty, since they were already part of the English language and accessible to LLMs. All scientific analogies contained explanations, which, again, we noted with the analogies.

In the next step, we used different LLMs to generate novel analogies (Claude Sonnet 4.5, ChatGPT 5, Gemini 2.5 and Mistral). The LLMs were instructed to adhere to the analogy syntax ("X is like a Y") and invent a source "Y" for a given "X" (the target). As a stimulus, we provide each LLM with the list of targets from the human analogies. We thus arrive at five distinct analogies per target (one human-authored and 4 LLM-generated). For the poetry dataset, we instructed the LLM to select two of our fixed set of 6 emotions and generate an analogy such that it conveys those emotions. The distribution of emotions in both

	Poetry			Science		
Origin	words	min	max	words	min	max
Human	20.4	9	56	54.0	19	110
LLM ∅	23.3	9	39	44.1	26	75
Claude	25.2	22	30	47.2	39	60
ChatGPT	14.8	9	19	32.6	27	41
Gemini	31.1	25	39	58.6	36	75
Mistral	21.9	16	27	37.8	26	50

Table 1: Analogy word count by model

human-written and LLM-generated metaphors leans toward "longing" and "love" (in 50% of the analogies), with the remaining emotions distributed roughly evenly.

For the science dataset, we provided identical instructions for syntax and novelty, omitted emotion, and prompted the LLM to serve as a science communicator. Each LLM generated one metaphor for each of the 2x20 targets, yielding 100 metaphors per dataset (20 written by humans, 80 generated by LLMs). Table 1 shows the average, minimum, and maximum word count of the analogies. The poetry and science subsets differ not only in the emotion conveyed but also in structure: human poetic analogies are, on average, 20.4 tokens long, whereas scientific metaphors are 54 tokens long, with much more variance in the human analogies. For LLMs, poetic analogies are longer than human ones (+3 tokens), whereas scientific analogies are shorter (-7 tokens), with substantial variance across LLMs.

4. Annotation and Analysis

We first describe our crowdsourced annotations. Then we conduct an initial analysis of features that distinguish human metaphors and features that make annotators *perceive* them as human-written.

4.1. Annotation via Mechanical Turk

We used Amazon Mechanical Turk[2] to crowdsource annotations along the following dimensions for all metaphors: quality (asking whether the analogy is a well-fitting image), writing (whether the author appears to be a professional writer), creativity, and comprehensibility, each rated on a Likert scale from 1 (very negative) to 5 (very positive). Further, we asked participants whether they thought the analogy could have been generated by a machine, with 1 indicating it was definitely human and 5 indicating they definitely considered it machine-generated.

For the poetic metaphors, we also asked participants to assign the analogies' conveyed emotions by multiple choice. For the science metaphors, we asked instead whether they found the metaphors helpful ("helpfulness") and whether they considered the author an expert on the scientific topic ("expertise"). We will provide the full set of questions in the Appendix. The results are shown in Table 2.

Each analogy was presented to 10 different annotators, yielding high standard deviations across all dimensions (0.74 for poetry and 0.78 for science, with "machine" showing the highest at 0.99). The overall picture indicates that distinguishing human-written analogies from generated analogies is difficult using the given dimensions, with average LLM scores close to human performance. Additionally, humans cannot identify LLM metaphors (as shown by the *machine* rating, which rates ChatGPT's analogies as most "human-like").

4.2. What marks human analogies?

Overall, we find very few dimensions that distinguish human-written from machine-generated analogies, and no significant differences for scientific metaphors.

For the poetry dataset, humans showed *lower* average scores for creativity, professional writing style, and overall quality. Across individual models, only Claude differed significantly from human writing in creativity and overall quality, whereas all models exceeded human scores for writing style. While both writing style and overall quality might have been influenced by translating and rephrasing the metaphors as analogies, this should not affect creativity. We also noted that annotators disagreed more when judging human poetry: we observed significantly lower inter-annotator agreement for writing style in human metaphors (standard deviation increased from

[2] https://www.mturk.com

Origin	quality		writing		creativity		comprehensible		machine?		helpful	expert
	Poe	Sci	Poe	Sci	Poe	Sci	Poe	Sci	Poe	Sci	(Sci only)	
Human	3.74	3.83	3.57	3.88	3.75	3.69	3.83	4.03	3.23	3.21	3.81	3.36
LLM ∅	**3.87**	3.87	**3.85**	3.85	**3.94**	3.75	3.86	4.02	3.25	3.24	3.73	3.36
Claude	**3.94**	3.93	**3.90**	3.92	**4.09**	3.82	3.83	4.06	3.24	3.14	3.78	3.34
ChatGPT	3.80	3.81	**3.73**	3.80	3.85	3.79	3.92	3.96	3.18	3.25	3.65	3.35
Gemini	3.87	3.90	**3.88**	3.83	3.87	3.73	3.78	3.99	3.33	3.33	3.82	3.40
Mistral	3.87	3.83	**3.89**	3.85	3.94	3.66	3.92	4.08	3.27	3.23	3.65	3.33

Table 2: Annotation scores for Poe[try] and Sci[ence]; scores ranging from 1 to 5; Values in boldface differ significantly ($p < 0.05$) from human analogies.

Model	Fear	Longing	Love	Pain	Pleasure	Regret	Overall
Human	73.5	57.0	68.5	63.5	73.5	76.0	68.7
LLM ∅	71.1	57.6	**54.1**	65.8	**65.5**	73.6	**64.6**
Claude	73.5	59.0	**56.0**	67.0	65.0	73.5	65.7
ChatGPT	69.5	61.5	**57.5**	68.5	65.0	74.0	66.0
Gemini	74.0	62.0	**56.0**	70.0	70.5	75.5	68.0
Mistral	67.5	48.0	**47.0**	57.5	**61.5**	71.5	**58.8**

Table 3: Annotators' emotion guessing accuracy by model (Poetry dataset only)

0.66 in LLMs to 0.82). While polished writing is a sign of "LLM language," human-authored poetic metaphors are more controversial, mirroring debates over the quality of artwork.

Subjectivity plays another role in differentiating humans and LLMs: humans use metaphors more reliably to convey emotion. Table 3 shows the accuracy with which annotators guessed the emotions underlying the analogies. Note that the task is different for human metaphors: LLMs selected the emotions for which they generated analogies. For humans, emotions were annotated post-hoc. Thus, we would expect the emotions in the LLM metaphors to be *easier* to infer because the analogies are specifically generated to convey those emotions, whereas in human analogies we measure rater agreement with expert annotators. Nevertheless, we find that, especially for "love" but also for "pleasure", conveying emotions is significantly more effective with human metaphors.

4.3. Which analogies *appear* human?

After having shown which features distinguish human metaphors, we now investigate what makes metaphors *appear* more human, i.e., which features correlate with a low "machine" rating in our annotation scheme. A "machine" score of 1 indicated that the annotator was certain the analogy was human-written, whereas 5 indicated that they attributed it to an LLM.

We find that human annotators cannot identify LLM-generated metaphors. This is already evident from the insignificant difference in the "machine" rating of human and LLM metaphors (0.02 for poetry, 0.03 for science). We also evaluated the accuracy of the annotators: by averaging each example's "machine" ratings and treating an average score below 3 as indicating human-written text, the accuracy of human annotation is below chance (0.42 for poetry, 0.41 for science). Still, the perception of what distinguishes human metaphors is partially accurate: as shown in Table 4, annotators assigned higher machine scores to poetic texts when they also rated them as more professionally written, which is consistent with the true correlation. However, ratings are inconsistent with respect to overall quality: for poetic metaphors, the annotators considered higher overall quality a sign of human-written language, the opposite of the actual result.

We also observe several differences between the two datasets: More comprehensible poetry appears more machine-like (perhaps because humans are assumed to ex-

Dimension	Poetry	Science
writing	+ **0.29**	+ 0.09
comprehensibility	+ **0.12**	− **0.13**
creativity	0.00	+ **0.26**
quality	+ **0.25**	+ 0.06
helpfulness	—	+ 0.02
expertise	—	+ **0.29**
emotion conveyance	− **0.26**	—

Table 4: Correlation (pearson's ρ) between ratings and "machine" score; boldface indicates significance (p<0.05). Positive correlation means a higher score makes the analogy look more "machine-like", negative means more "human-like".

ercise greater artistic freedom), and more comprehensible scientific metaphors appear more human-like to the annotators. For scientific metaphors, creativity and expertise are strongly correlated with the "machine" score; annotators attribute less creative and less informed texts to humans.

For poetry, a key feature distinguishing human metaphors relates to our previous assessment of conveying emotions: In cases where annotators could reliably detect the emotions of an analogy, they rated it as much more human-like ("emotion conveyance" in Tab. 4).

5. Detecting LLM Analogies

We now complement our analysis of human-specific analogy features with a series of classification experiments. We assess whether machine-generated metaphors can be identified solely from their text and whether our annotated features can serve as a valid proxy for identifying machine-generated content. We first distinguish human-authored analogies from generic LLM-generated metaphors, and then automatically classify human authors vs. individual models. In a final experiment, we test whether commercial large language models can tell machine-generated figurative language from human-written metaphors.

5.1. Classifiers and Baselines

We implement classifiers based on either textual features or our annotations. All algorithms were implemented using scikit-learn.

Classification by textual features

We first examine whether LLM-generated metaphors can be identified directly from their text. For this purpose, we extract sentence embeddings for each analogy using RoBERTa (Liu et al., 2019). We then classify the resulting vectors using a support vector machine (SVM) with leaving-one-out cross-validation (LOOCV). Note that this approach involves features generated by a language model, but does not use the LLM for direct classification.

Classification by annotation dimensions

We evaluate the predictive power of the annotated dimensions by training a classifier on the numerical annotated features and omitting the emotion labels. For each example, all annotation scores are averaged. Complementing our earlier analysis of individual correlations, this shows how our annotations interact for prediction. We evaluated multiple classification algorithms using LOOCV. We achieved the best results for poetry with logistic regression, and with Gradient Boosting for science.

Baselines

We provide two baselines for our classifiers: First, we use an informed baseline that uses text length as its only feature. As shown in Table 1, the length difference between human-written and machine-generated analogies is significant. We thus train another support vector machine with the number of tokens as the only feature and report it as the "Length" baseline. Second, we provide the performance of the human annotators on the balanced dataset. As described before (Sec. 4.3), we average the "machine" scores for each example and take a score below 3.0 (the midpoint of our Likert scale) to indicate a human author.

5.2. Humans vs. mixed LLMs

To generically distinguish human-written text from automatically generated metaphors, we downsample our dataset to achieve a balanced setup: we include all human metaphors (20 from poetry, 20 from science communication) and sample 20 machine-generated metaphors per dataset, randomly selecting 5

	Poetry	Science
Text	**0.83**	0.60
Annotation	0.70	0.60
Length	0.73	**0.70**
Human	0.45	0.60

Table 5: Classification accuracy for identifying machine-generated analogies, along with a naive baseline and the human performance.

per language model. The accuracy of a random baseline for this balanced dataset is 0.50.

Table 5 shows the results. For poetic metaphors, the text-based classifier achieves the highest accuracy (0.83), suggesting that substantial differences between LLM-generated text and human-authored metaphors can be derived from shallow textual features. The resulting accuracy exceeds the length baseline by 0.1. Annotation-based classification yields low accuracy, close to the length baseline (0.70 vs. 0.73).

For the science dataset, the results confirm our earlier annotation-based insights: for those more explanatory and less emotional metaphors, there is little discriminative signal for either human raters or model-based classifiers to exploit. Taking text length as the only feature is the most reliable way to distinguish LLM-generated scientific metaphors from human-written ones.

Human annotators appear to perform better on the science dataset than on the poetic metaphors, which is inconsistent with our other findings. Note that the downsampled dataset is insufficiently reliable for precise evaluation of human performance; as we have shown previously, accuracy on the complete data is 0.42 for poetry and 0.41 for science (cf. Sec. 4.3), which is equally low for both datasets.

5.3. Humans vs. individual LLMs

Because each LLM has its own specific language characteristics, we also try to distinguish human poetry from individual LLMs. For each LLM, we combine all analogies generated by that LLM and all human analogies, and then evaluate classification accuracy.

The results (Table 6) present a different picture than in our previous experiment: while, on average, the poetry dataset still yields better LLM identification, the results vary much more in the science context (with ChatGPT being easier to identify as an LLM than in the poetic context). This clearly shows that there is no consistent LLM language, and while no LLM can generate human-like poetic metaphors, some are excellent at generating human-like scientific analogies (Gemini, Mistral). ChatGPT and Claude might exhibit a particular "LLM syntax" (cf. Zamaraeva et al., 2025), which we tried to avoid by standardizing the analogy syntax. However, the scientific analogies often consist of multiple sentences and thus exhibit greater grammatical variation.

The length-based baseline often outperforms other approaches, except for Mistral's and ChatGPT's poetic analogies. Note that Mistral, which is easiest to detect as a machine for poetry by a large margin, was also the worst at conveying emotions. In science, only Claude's analogies slightly exceeded the length baseline. Overall, the recognition of most models appears to be possible by using surface patterns derived from text, but those patterns are not intuitively perceived as "machine-like" by humans. Furthermore, the models vary substantially in their ability to generate human-like text, both across models and across contexts.

5.4. Zero-Shot Classification with LLMs

Finally, we conduct an automatic annotation experiment using zero-shot prompting on the same Chatbots we used for text generation. We use the same balanced dataset from the first classification experiment (with mixed LLMs) and ask the same LLMs to predict whether each metaphor was written by a human or generated by a machine. To avoid data leakage, we used new accounts and different computers to retrieve the annotations.

The results (Table 7) show that, overall, these automatic annotators outperform both human raters and direct classification on both datasets. In particular, Claude and Gemini achieve perfect accuracy in detecting the origin of poetic metaphors, closely followed by ChatGPT. Mistral does not demonstrate comparable performance and operates even below chance for poetic metaphors. While ac-

	Poetry				Science			
(Origin)	Text	Anno.	Length	Human	Text	Anno.	Length	Human
ChatGPT	0.73	0.48	0.63	0.50	0.85	0.58	0.88	0.63
Claude	0.78	0.68	0.93	0.53	0.75	0.43	0.63	0.55
Gemini	0.80	0.75	0.93	0.58	0.53	0.53	0.83	0.63
Mistral	0.95	0.68	0.63	0.55	0.68	0.50	0.70	0.58

Table 6: Classification results for identifying individual LLMs

Model	Accuracy	
(Classifier)	Poe	Sci
Claude	1.00	72.50
Gemini	1.00	87.50
Mistral	0.35	50.00
ChatGPT	92.50	65.00

Table 7: Zero-shot prompt test for LLM metaphor identification using Chatbots as classifiers

curacy drops for science metaphors, Claude and Gemini still perform remarkably well at distinguishing human-written metaphors from LLM-generated writing, outperforming all classification approaches and baselines.

These findings are consistent with our previous embedding-based classification results. As shown previously, RoBERTa representations encode information that, to some extent, distinguishes human and machine metaphors, with more consistent results for poetic analogies. Given that contemporary chatbots rely on substantially larger and more sophisticated language models, it is unsurprising that some achieve strong performance in origin detection. At the same time, the variation across models suggests that this capability is neither uniform nor domain-independent; rather, it depends on both the underlying architecture and the characteristics of the metaphor domain. Determining what features precisely they can exploit remains a subject for future work.

6. Conclusion

We analyzed the differences between human-written and machine-generated novel metaphors. We first assembled two datasets from poetry and science communication, including human-written examples and their matched LLM-generated counterparts. An extensive crowdsourcing annotation revealed that, for science metaphors, the origin was largely unidentifiable using individual ratings, whereas successful emotional conveyance distinguished human poetic metaphors. Further, the human metaphors are written in less standardized language. We also showed that automated methods for predicting origins outperform human inspection, and that LLMs perform best at distinguishing between human- and machine-generated metaphors.

While our research confirms prior work showing that LLMs express scientific metaphors as proficiently as humans, we show that the uniqueness of human language remains evident in emotional poetic expressions. While LLMs are excellent at generating figurative language to explain complex concepts, they struggle with generating less concrete, more sentimental text.

Overall, we believe that more research is necessary to understand how the cognition of novel human metaphors differs from the mechanical assembly of figurative machine language, and how both can (or cannot) be distinguished from standard language processing. In future studies, we aim to identify additional features that distinguish human figurative writing from LLM-generated text, while accounting for differences across LLMs and new domains. Further, we aim to apply interpretability methods to language models to better analyze the mathematical processes behind metaphor generation. Exploring how challenges with emotional language relate to other LLM shortcomings can help identify the strengths and weaknesses of automated writing relative to human creativity, thereby enhancing the explainability of current LLMs and supporting their responsible use.

7. References

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2026. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279.

Yuri Bizzoni and Shalom Lappin. 2019. The effect of context on metaphor paraphrase aptness judgments. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 165–175, Gothenburg, Sweden. Association for Computational Linguistics.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112 1:193–216.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Berin Doru, Christoph Maier, Johanna Sophie Busse, Thomas Lücke, Judith Schönhoff, Elena Enax-Krumova, Steffen Hessler, Maria Berger, and Marianne Tokic. 2025. Detecting artificial intelligence–generated versus human-written medical student essays: Semirandomized controlled study. *JMIR Med Educ*, 11:e62779.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. arXiv. Preprint.

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado. Association for Computational Linguistics.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57, Atlanta, Georgia. Association for Computational Linguistics.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 24(4):296–309.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rosé. 2015. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 384–392, Prague, Czech Republic. Association for Computational Linguistics.

Kaidi Jia and Rongsheng Li. 2024. MD-PK: Metaphor detection via prompt learning and knowledge distillation. Preprint.

Kaidi Jia, Yanxia Wu, Ming Liu, and Rongsheng Li. 2025. Curriculum-style data augmentation for llm-based metaphor detection. Preprint.

Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, pages 115–135, New York, NY, USA. Association for Computing Machinery.

Oliver Kramer. 2025. Conceptual metaphor theory as a prompting paradigm for large language models. Preprint.

George Lakoff and Mark Johnson. 2003. *Metaphors we live by*. University of Chicago Press, Chicago. Originally published: Chicago : University of Chicago Press, 1980.

Yujie Lin, Jingyao Liu, Yan Gao, Ante Wang, and Jinsong Su. 2024. A dual-perspective metaphor detection framework using large language models. Preprint.

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. *CoRR*, abs/1907.11692.

Sofia Lugli and Carlo Strapparava. 2024. Multimodal chain-of-thought prompting for metaphor generation. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 523–530, Pisa, Italy. CEUR Workshop Proceedings.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 4221–4227.

Arthur Neidlein, Philipp Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736.

William Orwig, Emma R. Edenbaum, Joshua D. Greene, and Daniel L. Schacter. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior*, 58(1):128–136.

Natalie Parde and Rodney D. Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1535–1540.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Porter and Edouard Machery. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. Preprint.

Sadat Shahriar, Navid Ayoobi, and Arjun Mukherjee. 2025. The erosion of LLM signatures: Can we still distinguish human and LLM-generated scientific ideas after iterative paraphrasing? In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1118–1126, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xincheng Shou, Xiaoxi Huang, and Wenlong Xi. 2024. Conceptual metaphor theory guides gans for generating metaphors and interpretations. *IEEE Access*, PP(99):1–1.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source -

target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. Preprint.

Shanshan Wang, Junchao Wu, Fengying Ye, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2025. Benchmarking the detection of LLMs-generated Modern Chinese poetry. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9533–9552, Suzhou, China. Association for Computational Linguistics.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing LLM-generated and human-authored news text using formal syntactic theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Vienna, Austria. Association for Computational Linguistics.

Omnia Zayed, John P. McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, OpenAccess Series in Informatics (OASIcs), pages 10:1–10:17.

A. The Analogy Dataset

We provide the full dataset, including MTurk annotations, under https://osf.io/3bgn4/overview?view_only=addb6107053149e387df54647a30f063. Additionally, we show the list of targets in table 8 (poetry) and table 9 (science), as well as the full analogy set with origins for both datasets in table 10 (poetry) and 14 (science).

#	Poetry targets
1	you
2	we
3	my hands
4	the heart that's had enough
5	we
6	you lie upon my heart
7	my soul is dancing
8	my breath feels
9	the tip of your hair
10	your love
11	time
12	my life
13	some people
14	your body and your cloth
15	my love and yours
16	love
17	you and I
18	your touch on my shoulder, meant to ease my loneliness,
19	my mouth
20	regret and youth

Table 8: Poetry metaphor targets.

#	Science targets
1	learning in a neural network
2	evolutionary adaptation
3	Higgs field and particle mass
4	black holes
5	a superconductor
6	quantum entanglement of photon pairs
7	metaphors in science communication
8	the vagus nerve
9	your body's cells
10	the brain's default mode network (DMN)
11	telomeres on chromosomes
12	dysplastic cells
13	entropy in thermodynamics
14	spacetime curvature
15	epigenetics
16	carbon dioxide accumulation in the atmosphere
17	editing genes with the CRISPR technology
18	quantum superposition
19	protein folding
20	the event horizon of a black hole in space

Table 9: Science analogy targets.

B. Prompts for Dataset Creation

The following prompts are used to generate the analogies using the targets from our human-written dataset part (we omit the actual targets below, but they were given in the prompt). Note that the naming of the "source" might be confusing for someone familiar with conceptual metaphor theory, where it should be, in fact, the "target" or "tenor".

Poetic analogies

```
You are a poet.
I have a List of concepts /
phrases (""sources“).
I want you to invent metaphorical analogies
("vehicles") describing the sources in
emotional language.
Please chose, for each individual metaphor,
two sentiments from the following list:
love, pleasure, longing, pain, fear, regret

Then generate the analogy.
The analogies should have the form
[source] is / are like [vehicle],
[potentially an explanation]

Example: "My head feels like a pounding bomb
 of fire soup. I will burst at any time,
 radiating heat and dark glowing thoughts."

In the example, I would provide "My head"
 (source).  You would first pick two emotions
 (e.g., pain, longing),then choose a metaphor
 that expresses each.
The analogies should be new metaphors;
they may not be found in Google, primary
literature, or even a dictionary.
Please verify that.

Be poetic and original.
I will give you a list with the sources along
with ids and a list of sentiments.
Please give back the list, adding
```

columns for the complete metaphor, the isolated vehicle, an explanation, and the emotions you picked.

Scientific analogies

```
You are a professional science communicator.
I give you a List of concepts / words ("sources"). I want you to invent metaphorical analogies ("vehicles") explaining the word in the style of a science communicator.
The analogies should have the form [source] is / are like [vehicle]: [explanation]
Please explain a specific aspect about the source.

Example: "Cybersecurity is like seatbelts in cars: Initially deemed unimportant and uncomfortable, it turns out to be essential.
Now they are required by governance, but nobody would buy a car without seatbelts anyway."

The analogies should be new metaphors; they may not be there on Google or even in a dictionary.
Please verify that. In that respect, my example was bad, because car metaphors are rather common.
You can add specifications to the source; e.g., if I give you "cybersecurity" as a source, you can choose "Cybersecurity in large corporations" as a source.
Please provide the analogy ("is like.../ or "are like...") and an explanation.
I will give you a list with the sources along with ids.
Please give back the list, adding columns for the complete metaphor and the vehicle.
```

C. Mechanical Turk Tasks

In the following, we provide instructions and questions used to elicit annotations from Mechanical Turk.

C.1. Poetry

Instructions: You are given a metaphor in form of an analogy. Please rate the metaphor for us. We will also ask you for the emotion this metaphor conveys, and for its origin.

Analogies are “is like” or “are like” sentences, such as “Kids are like glowworms: Both are bright, nervous, never sleep at night, and a wonder of nature.” Please tell us what you think about the analogy. There is no right or wrong, we want your opinion.

The heart that’s had enough is like a locked garden after the storm—its gates rusted shut, its roses still bleeding perfume into the wind.

1. Do you think it is a good, well-fitting analogy? 1 means it’s bad, 5 means it’s excellent. *(Score from 1 to 5)*
2. Do you think the writer was professional? 1 means it’s very likely an inexperienced writer or student, 5 means it’s an excellent professional writer. *(Score from 1 to 5)*
3. Do you think it is comprehensible? 1 means it’s very cryptic, even with explanation; 5 means it’s right on point and easy to understand, even without explanation. *(Score from 1 to 5)*
4. Do you think it is creative? 1 means it’s very common, 5 means it’s very original and unusual. *(Score from 1 to 5)*
5. Which emotions are described with the metaphor? You can check one or two. Please tick at least one, at most two.
   ☐ love ☐ pleasure ☐ longing ☐ fear ☐ regret ☐ pain
6. Do you think the analogy has been written by a machine? 1 means no, it’s definitely a human, 5 means yes, it’s definitely a machine. *(Score from 1 to 5)*
7. Space for comments if you want to tell us something: *(Free text field)*

Figure 1: Annotation interface for the poetry metaphor dataset as shown to crowd workers.

C.2. Science

Instructions: You are given a scientific metaphor in form of an analogy. Please rate the metaphor for us.

Analogies are “is like” or “are like” sentences, such as “Kids are like glowworms: Both are bright, nervous, never sleep at night, and a wonder of nature.” Please tell us what you think about the analogy. There is no right or wrong, we want your opinion. The analogies are all about science topics. You do not have to be proficient in the scientific topic to answer the question.

Editing genes with CRISPR is like using a GPS-guided scalpel: it navigates to an exact address in the genome and makes a precise cut, enabling repairs that previously required guesswork.

1. Do you think it is a good, well-fitting analogy? 1 means it’s bad, 5 means it’s excellent. *(Score from 1 to 5)*
2. Do you think the writer was professional at writing? 1 means it’s very likely an inexperienced writer or student, 5 means it’s an excellent professional science communicator. *(Score from 1 to 5)*
3. Do you think it is comprehensible? 1 means it’s very cryptic, even with explanation; 5 means it’s right on point and easy to understand, even without explanation. *(Score from 1 to 5)*
4. Is the metaphor helpful to understand the scientific concept? 1 means it’s just confusing, 5 means you learned something new just by reading this metaphor. *(Score from 1 to 5)*
5. Do you think it is creative? 1 means it’s very common, 5 means it’s very original and unusual. *(Score from 1 to 5)*
6. We need to check whether people read our questions — please just tick “3” for the next answer. *(Score from 1 to 5)*
7. Do you think the writer was an expert on the scientific topic? 1 means no, it seems he or she just writes about it, 5 means it’s probably a scientist working on the topic. *(Score from 1 to 5)*
8. Do you think the analogy has been written by a machine? 1 means no, it’s definitely a human, 5 means yes, it’s definitely a machine. *(Score from 1 to 5)*
9. Space for comments if you want to tell us something: *(Free text field)*

Figure 2: Annotation interface for the science analogy dataset as shown to crowd workers.

Table 10: Poetry metaphors and their origin.

#	Origin	Metaphor
1	Human	You are like a glass of water after too many beers.
2	Human	We are like a left-handed and a right-handed person who dream about flying: you have a wing on your left shoulder, and I do have one on my right, and we want to grow together and take off. Because we're afraid to tear each other apart, we just hold hands, and we scratch each others shoulder.
3	Human	My hands are a chopping block and I cannot touch him. I cannot touch him without not touching me.
4	Human	The heart that's had enough stays shut. Like an oyster that cloisters a spoil of pearls, Untouched.
5	Human	We were like an aeroplane made from a man and wife. We hovered a little above the earth. We even flew a little.
6	Human	You lie upon my heart as on a nest, Folded in peace, for you can never know How crushed I am with having you at rest Heavy upon my life.
7	Human	my soul is dancing like tripping shoes with breathless fire towards you
8	Human	my breath feels like sand and lightning of thorns
9	Human	The tip of your hair is like a scorpion tail lying on the moon of your face, an omen unveiled.
10	Human	Your love is like the sip of water given to a dove, just before the dagger touches its throat.
11	Human	Time is like unstoppable snow that settles on our brows and hair.
12	Human	My life without you is like an empty notebook book, which keeps turning pages after page in longing for your face.
13	Human	Some people are like silkworm that spend their whole life weaving a cage, yet dreaming only of flying.
14	Human	Your body and your cloth are like the Garden of Eden and the gate to its secret. As I open the gate, spring spills into the room, and in my arms, a flower named you unfolds.
15	Human	My love and yours, like two rivers, merged into one great flow, but alas! that great river ended in a swamp.
16	Human	Love is like a wanderer who enters the house without knocking.
17	Human	You and I are like the sun and the morning star. I fade away when you shine.
18	Human	Your touch on my shoulder, meant to ease my loneliness, is like brushing snow from a snowman's shoulder.
19	Human	My mouth is like a scar on my face. I shut my lips so tightly as if the scar is healing in silence.
20	Human	regret and youth are like thorn and flower. Only thorns remain after a flower dies.
21	ChatGPT	You are like a candle burning in a locked room, filling the air with warmth I can't escape.
22	ChatGPT	We are like two moths circling the same flame, thrilled by the light and terrified of its heat.
23	ChatGPT	My hands are like wilted leaves still trembling in the wind, remembering the storms they once held.
24	ChatGPT	The heart that's had enough is like a fortress whose gates rust shut, still echoing with old cries.
25	ChatGPT	We are like waves meeting in the dark, touching only long enough to shimmer and vanish.

Table 11: Poetry metaphors and their origin. (continued)

#	Origin	Metaphor
26	ChatGPT	You lie upon my heart like a velvet stone—soft in weight, heavy in silence.
27	ChatGPT	My soul is dancing like a flame in a mirror, forever chasing its own reflection.
28	ChatGPT	My breath feels like the ghost of a storm that never reached the shore.
29	ChatGPT	The tip of your hair is like the edge of dawn—quiet light brushing away my night.
30	ChatGPT	Your love is like a house built on tides—beautiful, shifting, impossible to live in.
31	ChatGPT	Time is like a patient thief, stealing faces first, then memories.
32	ChatGPT	My life is like a field of echoes, each joy calling softly for another.
33	ChatGPT	Some people are like unplayed songs, left waiting in the dust of the piano.
34	ChatGPT	Your body and your cloth are like twin rivers of silk, flowing through my dreams.
35	ChatGPT	My love and yours are like two mirrors facing each other—endless, but never touching.
36	ChatGPT	Love is like a wild fruit—sweet at first bite, bleeding on the tongue.
37	ChatGPT	You and I are like parallel shadows stretching toward a sunset that will not wait.
38	ChatGPT	Your touch on my shoulder, meant to ease my loneliness, is like sunlight through broken glass—warm, but it cuts.
39	ChatGPT	My mouth is like a gate trembling before confession.
40	ChatGPT	Regret and youth are like twin ghosts holding hands, haunting the same photograph.
41	Claude	You are like a lighthouse built on dissolving sand, pulling my gaze across impossible distances while the ground beneath your foundation whispers of collapse.
42	Claude	We are like two mirrors facing inward, creating an infinite tunnel of reflected joy that we can never step inside without shattering the glass.
43	Claude	My hands are like empty theaters after the last performance, still warm with phantom applause, aching for actors who will never return to the stage.
44	Claude	The heart that's had enough is like a well that's learned to fear the bucket, flinching at the sound of rope uncoiling in the dark.
45	Claude	We are like two flames sharing the same candle wick, burning twice as bright while knowing we consume our foundation faster together.
46	Claude	You lie upon my heart like snow on a greenhouse roof, beautiful in your cold weight, keeping darkness in while promising to melt with spring that may never arrive.
47	Claude	My soul is dancing like a kite with a severed string, delirious with newfound freedom yet falling slowly toward the inevitable earth.
48	Claude	My breath feels like water trying to climb stairs, struggling upward against its nature, spilling and gathering, never quite reaching the landing.
49	Claude	The tip of your hair is like the last page of an unfinished book, trailing off mid-sentence, inviting my fingers to write the ending that doesn't exist.
50	Claude	Your love is like an ancient bridge I've crossed too heavily, hearing new groans in the timber with each return journey, wondering which crossing will be my last.

Table 12: Poetry metaphors and their origin. (continued)

#	Origin	Metaphor
51	Claude	Time is like a river carving its canyon backward, eroding yesterday before tomorrow, leaving me stranded on a shrinking island of present moments.
52	Claude	My life is like a garden where I've planted only question marks, now blooming into shapes I can't name, beautiful and useless for harvest.
53	Claude	Some people are like photographs developing in reverse, growing dimmer with each year of knowing them, until they're just blank paper that once held faces.
54	Claude	Your body and your cloth are like a violin still singing after the bow has lifted, the fabric holding the vibration of your form in its folds.
55	Claude	My love and yours are like two roots growing from opposite sides of a stone, reaching toward each other through the dark, scraping granite with every inch of progress.
56	Claude	Love is like a bird building its nest inside a closing fist, singing sweetly while arranging twigs in a space that's learning to crush.
57	Claude	You and I are like two clocks set to different time zones, both keeping perfect time, both correct, never sharing the same hour together.
58	Claude	Your touch on my shoulder, meant to ease my loneliness, is like placing a single candle in a cathedral of darkness—it only shows me how vast the emptiness truly is.
59	Claude	My mouth is like a train station for words that missed their departure, crowding the platform with tickets to conversations we never boarded.
60	Claude	Regret and youth are like twins born facing opposite directions, one watching the sunrise the other never saw, the other seeing the sunset the first one can't remember.
61	Gemini	You are like a newly discovered planet, veiled in a brilliant magnetic storm. The light is irresistible, but the turbulence threatens to shred my vessel.
62	Gemini	We are like a half-forgotten song, played on a scratched vinyl record, where the melody still thrills, but the skips are where the joy falls silent.
63	Gemini	My hands are like the dry riverbeds of an ancient drought, cracked and hollowed out from where the water used to run, yearning for the returning tide.
64	Gemini	The heart that's had enough is like a fortress whose drawbridge is rusted shut and reinforced with diamond. It is impenetrable, guarding against invasion, but also trapping its own soft, glowing core inside.
65	Gemini	We are like a shared glance across a crowded, soundless cathedral, a fleeting moment of pure, vibrant recognition that asks for more, but cannot break the distance between the aisles.
66	Gemini	You lie upon my heart like the heavy, cold marble slab of a tombstone, pressing out my breath and sealing the space where lighter, living feelings should stir.
67	Gemini	My soul is dancing like a sudden, dazzling solar flare blooming from a star, pure energy unleashed, bathing the dark silence in golden, unburdened heat.
68	Gemini	My breath feels like a fraying rope pulled taut across a sharpened rock. Each shallow inhalation cuts deeper, and the slightest movement will snap the whole fragile lifeline.
69	Gemini	The tip of your hair is like the finest strand of spun moonlight, impossibly delicate and beautiful, a silver thread I follow through the dark, knowing I can never gather it all.
70	Gemini	Your love is like a massive, silent glacier, floating under a summer sky. It is beautiful and cool to the touch, but the unseen, submerged power of it holds the threat of inevitable, indifferent collapse.
71	Gemini	Time is like a swift, unseen river that only flows backward, dragging all our moments into a dark wake, where every current regret can be seen but never retrieved.
72	Gemini	My life is like an enormous, half-completed canvas of storm-tossed seas, filled with rich, deep colors, but the horizon is yet unpainted, waiting for the one shade that will make it whole.
73	Gemini	Some people are like a flock of mirrored crows that never blink, watching your every movement with an unsettling, total clarity, showing you the exact shape of your past mistakes.
74	Gemini	Your body and your cloth are like a newly bloomed midnight-blue flower that folds inward when touched. Its textures are a soft pleasure, and its modesty is a precious invitation to reverence.

Table 13: Poetry metaphors and their origin. (continued)

#	Origin	Metaphor
75	Gemini	My love and yours are like the twin, opposing currents of a subterranean spring, both necessary for the flow of life, but their ceaseless collision creates a powerful, aching pressure in the stone chambers.
76	Gemini	Love is like a vast, empty ballroom with a single, spinning chandelier, shedding light on the dust, desirable for its brilliance, but terrifying in the echoing silence it highlights.
77	Gemini	You and I are like a perfectly tuned, high-tension wire stretched between two distant peaks. The vibration when we touch is pure music, but the stretch of the cable represents an endless, unyielding distance.
78	Gemini	Your touch on my shoulder, meant to ease my loneliness, is like a sudden, brief rain shower on a parched battlefield. It offers a gentle love, yet only deepens the regret for the drought I allowed to settle there.
79	Gemini	My mouth is like a hollow made of warm, sculpted porcelain, sensitive to the lightest press and capable of soft beauty, but utterly fragile and fearful of the noise that could crack its gentle shell.
80	Gemini	Regret and youth are like two children playing with a shattered, silver mirror. They are forever trying to piece together a beautiful reflection from the painful, sharp fragments of what they once saw, wishing the surface were whole again.
81	Mistral	You are like a constellation I trace with my fingertips in the dark—each star a promise, each silence a galaxy of what could be.
82	Mistral	We are like two rivers that forget they are bound for the same sea, carving separate canyons just to meet again in the roar of the falls.
83	Mistral	My hands are like birds that built a nest in your absence, only to find their wings too heavy with the weight of waiting.
84	Mistral	The heart that's had enough is like a locked garden after the storm—its gates rusted shut, its roses still bleeding perfume into the wind.
85	Mistral	We are like a song played backward—each note a memory, each silence a place we dared not go.
86	Mistral	You lie upon my heart like a map folded wrong—every crease a border I can't cross, every edge a place I've already lost.
87	Mistral	My soul is dancing like embers leaping from a fire—each spark a laugh, each fall a whisper of ash.
88	Mistral	My breath feels like a kite string tangled in the branches of a tree I can't climb—each pull a gasp, each knot a secret I can't name.
89	Mistral	The tip of your hair is like the last brushstroke on a painting I never finished—soft, deliberate, and already fading.
90	Mistral	Your love is like a lighthouse I mistook for the moon—its beam a guide, its absence a tide that pulls me under.
91	Mistral	Time is like a thief who only steals what you've already lost—each pocket empty, each memory a ghost in the hallway.
92	Mistral	My life is like a book with every other page torn out—each remaining word a clue, each missing one a wound.
93	Mistral	Some people are like shadows cast by a fire—warm to look at, but impossible to hold.
94	Mistral	Your body and your cloth are like a letter written in ink that runs when touched—each word a promise, each smear a secret.
95	Mistral	My love and yours are like two mirrors facing each other—each reflection a world, each crack a place the light escapes.
96	Mistral	Love is like a bridge built between cliffs—each step a risk, each handrail a story we tell ourselves to keep walking.
97	Mistral	You and I are like two clocks in the same room—each tick a heartbeat, each chime a place we almost met.
98	Mistral	Your touch on my shoulder is like a key turned in a lock that opens into another lock—each click a hope, each stuck door a silence.
99	Mistral	My mouth is like a cave where echoes go to die—each word a ghost, each silence a tomb.
100	Mistral	Regret and youth are like twin rivers—one clear and rushing, the other thick with silt, both emptying into the same dark sea.

Table 14: Science analogies and their origin.

#	Origin	Metaphor
1	Human	Learning in a neural network is like sculpting with a million tiny hands. Each hand makes a minuscule adjustment to the clay (the data), and through countless iterations, a complex form (the learned pattern) emerges, far beyond what any single hand could create.
2	Human	Evolutionary adaptation is like a tinkerer constantly modifying an existing machine, rather than a master engineer designing from scratch. Each modification has to work with what's already there, leading to sometimes clumsy but functional solutions.
3	Human	The Higgs field is like a celebrity party, and particles are the guests. Some particles are very popular and are constantly mobbed by the crowd, making them move slowly (giving them mass). Others are less popular and can zip through the room unhindered (having less or no mass).
4	Human	A supermassive black hole at the center of a galaxy is like the quiet, gravitational heart of a cosmic city. It doesn't actively 'eat' everything, but its immense gravity organizes the flow of stars and gas around it, defining the very structure and dynamics of the entire metropolis.
5	Human	A room-temperature superconductor is like a dance floor where a rowdy conga line suddenly becomes an orderly ballroom dance.
6	Human	Quantum Entanglement of photon pairs is like a pair of shows kept in two separate boxes The moment you identify one shoe, the nature of the other (whether it is the left or right shoe) is instantly discerned, regardless of its location in the universe. However, the intriguing factor is the inherent uncertainty associated with the identification process until the exact moment of observation.
7	Human	Metaphors [in science communication] are like zealous fungi that colonise different ecological niches, their presence and impact proliferates across key biological concepts.
8	Human	The vagus nerve is like the body's internal internet cable, running from the brainstem down to the gut and major organs. It's the primary two-way highway for unconscious communication, sending signals about our internal state up to the brain and sending regulatory commands back down.
9	Human	Your body's cells are like meticulously managed mini-cities, each with its own power plants (mitochondria), waste disposal systems (lysosomes), communication networks, and factories (ribosomes) constantly working in coordinated harmony to sustain the larger organism.
10	Human	The brain's default mode network (DMN) is like the mind's internal radio station that plays when you're not actively listening to anything else. It's where your mind wanders, daydreams, and reflects on self and others, but sometimes, in conditions like depression, it can get stuck on a repetitive, negative playlist.
11	Human	Telomeres on chromosomes are like shoelace aglets: Telomeres are the ends of chromosomes, and much like the cap at the tip of your shoelace (an aglet), they help maintain chromosome integrity by preventing the ends from fraying. As cells divide, telomeres shorten like worn aglets.
12	Human	Dysplastic cells are like weeds that have overgrown a garden. They choke everything else out, explaining pancytopenia (low blood counts). The way to treat it is to use intensive chemotherapy - a 'weed killer' that clears out the abnormal cells so normal cells can grow back.
13	Human	Entropy in thermodynamics is like a teenager's meyys bedroom If no energy or work is put in, a room quickly becomes messy and disordered - high entropy. If energy is input in the form of cleaning up and putting everything away, the room returns to a state of order or low entropy. The universe tends toward disorder unless energy is expended to maintain organization.

Table 15: Science analogies and their origin. (continued)

#	Origin	Metaphor
14	Human	Spacetime curvature is like a bowling ball on a rubber sheet: Imagine a rubber sheet stretched out. Place a bowling ball on the sheet - the material deforms around the mass. Roll a golf ball across the sheet and its motion changes in response to the deformation. This is how mass curves spacetime and creates gravity - not as a mysterious force, but as objects following curved paths in warped space.
15	Human	Epigenetics is like a pianist playing the piano: If the genome is the piano keyboard, specific and well defined, then epigenetics is the pianist who plays it. DNA provides the static blueprint, but epigenetic processes determine when and how each key can be struck, changing the tune being played. The piano remains important, but the pianist - the epigenome - brings the static genetic code to life.
16	Human	Carbon Dioxide accumulation in the atmosphere is like a bathtub filling with water: Pouring water in is like emissions; the drain is Earth's ability to absorb CO2. If you fill it faster than it drains, the level rises. We are now adding about twice as much CO2 as natural processes remove. Stabilizing requires cutting emissions to the rate at which they are removed - at least half.
17	Human	Editing genes with the CRISPR technology is like using malware on a city's internet: The greater metropolis represents the body, the suburbs are organs, the buildings are cells, the people are proteins, and the internet is DNA. CRISPR is malware that can search for any chosen 20-character line of code and corrupt it.
18	Human	Quantum superposition is like a marble rolling in a bowl: In atomic ion experiments, we can make an atomic marble roll back and forth just as a marble would in a real bowl. At some instance the atom will be on the right side, then on the left side. But we can also create a state where the atom is both on the right side of the bowl and the left side of the bowl at the same time.
19	Human	Protein folding is like Origami: Proteins, the essential building blocks of life, can only function if they are folded properly. The protein sheet is similar to origami paper. Both need to be measured and folded accurately. Precision, perfection and patience are of utmost importance. If we make a mistake in one of the folds, we end up with an unstable, inaccurate structure, just as misfolded proteins can lead to disease.
20	Human	The event horizon of a black hole in space is like a waterfall: Space flows toward the black hole like a river flows toward a waterfall. Outside the horizon, space falls inward at leass than the speed of light, so a "photon-fish" swimming upstream can make headway against the current. At the horizon, space falls at exactly the speed of light - a photon swimming directly upstream stays in place, swimming frantically but going nowhere. Inside the horizon, the space waterfall falls faster than light, carrying everything inevitable toward the singularity. Just as a fish sweot opver a waterfall cannot swim back up, nothing past the even horizon can return.
21	ChatGPT	Learning in a neural network is like a coral reef building itself polyp by polyp: each training example lays down a wafer-thin adjustment, and over countless iterations the once-open water hardens into an intricate architecture that channels future currents of information.
22	ChatGPT	Evolutionary adaptation is like a jazz standard being re-interpreted by successive generations of musicians: every player keeps the riffs that groove with the environment and drops the sour notes, gradually refining the tune for its changing audience.
23	ChatGPT	The Higgs field and particle mass are like invisible velvety curtains hanging across a stage: when different actors push through, the fabric clings to some more than others, slowing them just enough to give them the sensation of weight.
24	ChatGPT	Black holes are like a library's incinerator chute with a one-way, time-locked door: once a book slides in it can never be borrowed again, yet the furnace glows with a cryptic heat that hints a whole story was sacrificed.
25	ChatGPT	A superconductor is like a circular maglev race track coated in quantum Teflon: nudge the runners once and they loop forever, paying no energy toll to friction or resistance.

Table 16: Science analogies and their origin. (continued)

#	Origin	Metaphor
26	ChatGPT	Quantum entanglement of photon pairs is like twin improvisational dancers on opposite stages wearing motion-capture suits: spin one and, regardless of distance, the other instantly mirrors the move, as if choreography itself ignores geography.
27	ChatGPT	Metaphors in science communication are like inflatable hiking bridges stretched over conceptual canyons: they may wobble, but they let audiences cross gulfs of unfamiliar math without waiting for a permanent suspension bridge.
28	ChatGPT	The vagus nerve is like a diplomatic courier on a slow river barge that threads every organ kingdom: it delivers cease-fire telegrams that persuade local militias of muscle and gut to stand down from fight-or-flight hysteria.
29	ChatGPT	Your body's cells are like billions of pocket-sized startups all reading the same business plan: each pivots to a niche market, collaborates with neighbors, and gracefully shutters when conditions demand a reboot.
30	ChatGPT	The brain's default mode network is like a city's nocturnal street-sweeper fleet: as rush-hour thinking eases, they glide through mental avenues, tidying memories and sketching tomorrow's routes while the mind idles at the lights.
31	ChatGPT	Telomeres on chromosomes are like wax seals on old letters: they protect the message inside from fraying or contamination, but each time the letter is reopened and re-sealed, a bit of that protective wax gets used up.
32	ChatGPT	Dysplastic cells are like architectural blueprints with smudged lines: the overall shape is still recognizable, but the distortions warn you that the construction might soon go dangerously off-plan.
33	ChatGPT	Entropy in thermodynamics is like a library whose books slowly migrate out of order: without constant effort to re-shelve them, the natural tendency is toward increasing disorganization.
34	ChatGPT	Spacetime curvature is like a riverbed carved by passing water: the flow (mass–energy) shapes the channel, and in turn the channel guides every future current that moves through it.
35	ChatGPT	Epigenetics is like sticky notes placed on a manuscript: they don't change the words themselves, but they tell the reader which passages to highlight, whisper, skip, or shout aloud.
36	ChatGPT	Carbon dioxide accumulation in the atmosphere is like a slowly thickening quilt over a cooling body: at first it feels comforting, but add enough layers and the trapped heat becomes oppressive.
37	ChatGPT	Editing genes with CRISPR is like using a GPS-guided scalpel: it navigates to an exact address in the genome and makes a precise cut, enabling repairs that previously required guesswork.
38	ChatGPT	Quantum superposition is like a coin humming on its edge before it falls: it occupies many potential outcomes at once, only settling into a definite state when it finally tips.
39	ChatGPT	Protein folding is like origami instructions hidden inside the paper itself: the shape emerges not from an external folder, but from the internal cues that make the paper fold along certain lines.

Table 17: Science analogies and their origin. (continued)

#	Origin	Metaphor
40	ChatGPT	The event horizon of a black hole is like a cosmic one-way security turnstile: once anything passes through, even information loses any legal right to exit again.
41	Claude	Learning in a neural network is like a jazz ensemble improvising together: Each musician (neuron) listens to subtle cues from others, gradually synchronizing their performance through countless micro-adjustments until the entire group creates harmonious music none could produce alone.
42	Claude	Evolutionary adaptation is like a master locksmith working in complete darkness: With no blueprint or plan, countless random key variations are tried against an ever-changing lock (environment), with only the keys that happen to fit being copied and modified further.
43	Claude	The Higgs field is like an invisible cosmic molasses that permeates all space: Most particles get "stuck" in this field to varying degrees, which we perceive as their mass, while photons slip through completely unstuck, remaining massless and traveling at light speed.
44	Claude	Black holes are like cosmic trapdoors with broken hinges: Once matter falls through the event horizon, there's no mechanism in the universe's physics that can swing that door back open—not because the door is locked, but because the hinges themselves have been fundamentally severed by extreme spacetime curvature.
45	Claude	A superconductor is like a perfectly choreographed conga line of electrons: When cooled sufficiently, electrons pair up and move in perfect synchronization with zero resistance, like dancers who've practiced together so extensively they never step on each other's feet or break the rhythm.
46	Claude	Quantum entanglement of photon pairs is like identical twins sharing the same dream: No matter how far apart they are, when one twin experiences something in the dream, the other instantly experiences the complementary sensation—not through any signal between them, but because they're fundamentally sharing the same quantum "dream state."
47	Claude	Metaphors in science communication are like cognitive bridges built from familiar materials: They allow minds to cross from the known territory of everyday experience into the alien landscape of scientific concepts, but sometimes these bridges can only carry you so far before you need to learn the native language of mathematics and experimentation.
48	Claude	The vagus nerve is like the body's diplomatic ambassador: It travels extensively throughout the body's territories, constantly negotiating between the brain's central government and the various organ systems, sometimes calming tensions and other times relaying urgent messages that require immediate attention.
49	Claude	Your body's cells are like autonomous city-states in a vast federation: Each cell governs itself with its own internal politics (organelles), trades resources with neighbors, follows federal laws (genetic instructions), and can even vote to self-destruct for the good of the whole civilization when compromised.
50	Claude	The brain's default mode network is like a theater's backstage crew during intermission: When the main performance (focused attention) pauses, this network springs into action—reviewing past scenes, rehearsing future acts, and maintaining the narrative continuity that makes the overall production of consciousness feel seamless.
51	Claude	Telomeres on chromosomes are like the ink cartridges in a printer: each print job consumes a bit of ink, and while the printer keeps working fine, eventually the cartridge runs dry and the machine simply stops printing altogether — not because anything is broken, but because it has consumed its finite supply.
52	Claude	Dysplastic cells are like employees trained by a mentor who developed bad habits: they're not yet rogue or malicious, but their work shows troubling deviations from proper protocols — still salvageable with retraining, but if left uncorrected, they may eventually go fully off-script.

Table 18: Science analogies and their origin. (continued)

#	Origin	Metaphor
53	Claude	Entropy in thermodynamics is like paint spilled on a floor: technically, all the molecules are still there, and the laws of physics don't forbid them from spontaneously re-collecting into the can — it's just so astronomically improbable that you'd wait longer than the universe's lifetime for it to happen.
54	Claude	Spacetime curvature near massive objects is like reading through a book where the text gets progressively denser: near the center, every inch of page contains far more words, so traversing the same physical distance takes much longer and bends your attention inward.
55	Claude	Epigenetics is like placing sticky notes throughout a reference manual: the printed text hasn't changed, but those flags determine which chapters actually get read and how often — and these notes can be moved, added, or removed based on changing circumstances.
56	Claude	Carbon dioxide accumulation in the atmosphere is like a sink where the drain is slowly clogging: even if you reduce how much water you're pouring in, the level keeps rising because less is going down than before — stabilization requires not just turning down the tap, but unclogging the drain or turning the flow down far more than most expect.
57	Claude	CRISPR gene editing is like giving a librarian GPS coordinates for a specific sentence in one book among millions, plus a bottle of correction fluid: she can find exactly that passage and white out a word — but what happens after depends on whether the library has a good replacement text or just guesses what should go there.
58	Claude	Quantum superposition is like a sealed ballot that hasn't been counted yet: it's not that we simply don't know the vote — the voter genuinely hasn't committed to either candidate until the moment the envelope is opened, at which point the choice crystallizes into one definite outcome.
59	Claude	Protein folding is like a parachute that must pack itself into its container: the final shape is utterly critical for function, the starting material is just loose fabric, and while the correct configuration exists, getting there reliably from a tangled state is an engineering feat that nature mastered long before we understood it.
60	Claude	The event horizon of a black hole is like the point on a descending escalator where you're moving downward faster than you could ever climb back up: there's no wall, no special marking, nothing you'd notice as you cross it — but from that step onward, every direction you walk leads inexorably to the bottom.
61	Gemini	Learning in a neural network is like a glacier's advance: It's a slow, imperceptible reshaping of the landscape, where each tiny movement of ice and rock, accumulated over vast stretches of time, ultimately carves out immense valleys and peaks, reflecting the persistent, subtle adjustments of synaptic weights that sculpt a network's understanding of data.
62	Gemini	A superconductor is like a perfectly choreographed ballet: All the dancers (electrons) move in absolute unison and harmony, anticipating each other's every move, flowing effortlessly across the stage (the material) without any collisions or wasted motion,
63	Gemini	The Higgs field and particle mass are like a persistent rumor in a bustling market: The more "popular" a particle (the more it interacts with the Higgs field), the more it gets "slowed down" and "weighed down" by the constant chatter and interactions with the rumor (the Higgs field particles), giving it inertia and making it harder to push around, while less popular particles sail through relatively unimpeded.

Table 19: Science analogies and their origin. (continued)

#	Origin	Metaphor
64	Gemini	Black holes are like cosmic tax collectors: They have an undeniable gravitational pull, drawing in everything within their jurisdiction (event horizon), irrespective of its nature – light, matter, information – accumulating it into an incredibly dense, unrecoverable "revenue stream," from which nothing can escape their ultimate collection.
65	Gemini	A superconductor is like a perfectly choreographed ballet: All the dancers (electrons) move in absolute unison and harmony, anticipating each other's every move, flowing effortlessly across the stage (the material) without any collisions or wasted motion, resulting in a performance (current) that is flawlessly efficient and entirely without friction.
66	Gemini	Quantum entanglement of photon pairs is like synchronized swimming for light: Two performers (photons) are linked in such a way that even when separated by vast distances, a precise measurement of one's orientation instantly determines the corresponding orientation of the other, as if they are performing a complex, pre-programmed routine perfectly in sync, no matter how far apart they are.
67	Gemini	Metaphors in science communication are like interstellar probes: They are carefully crafted vessels launched from the familiar shores of everyday experience into the vast, unknown cosmos of complex scientific concepts, carrying a simplified yet accurate representation that allows the mind to explore and grasp distant, otherwise inaccessible phenomena.
68	Gemini	The vagus nerve is like a diplomatic courier service for your body: It's a vast, bidirectional information highway, carrying urgent and nuanced messages of status and need between your brain (the central government) and most of your major organs (various departments), ensuring coordinated responses and maintaining internal harmony, even during times of stress.
69	Gemini	Your body's cells are like a city's micro-economies: Each cell is a self-contained unit, constantly engaging in intricate metabolic transactions, importing raw materials, generating energy currency (ATP), producing specialized goods and services (proteins, hormones), and managing waste, all while contributing to the overall function and prosperity of the larger urban organism.
70	Gemini	The brain's default mode network (DMN) is like an orchestra's tuning session: Even when the main performance (focused task) isn't happening, this specific ensemble of brain regions is subtly active, harmonizing its instruments (neural circuits), reflecting on past pieces, and anticipating future ones, ensuring that when the conductor's baton rises, the entire symphony is ready to play in perfect concert.
71	Gemini	Telomeres are like the burn rate of a scented candle: The full candle represents the chromosome with its cap, and the progressive melting and shortening of the candle's wick and wax is the cell division process. Once the candle burns down to the base, its function ends, much like the limit of cellular division (the Hayflick limit) when telomeres are too short.
72	Gemini	Dysplastic cells are like the pilot light on an old gas stove: They are not a full-blown inferno (cancer) yet, but they represent an abnormal state—an energy-inefficient, low-level ignition that should not be continuously on. It's a clear, persistent warning sign of instability that could lead to an uncontrolled, dangerous blaze if not addressed.
73	Gemini	Entropy in thermodynamics is like the diffusion of a drop of colored ink in a glass of still water: The highly ordered, concentrated drop of ink will inevitably spread out until the color is uniformly and irreversibly distributed throughout the entire volume. This movement from a state of low-probability order to a state of high-probability disorder is the universe's natural tendency toward maximum entropy.
74	Gemini	Spacetime curvature is like the weight-based distortion of a trampoline surface: A massive object, such as a planet or star, is not pulling smaller objects toward it with an invisible "force," but rather, its mass is warping the fabric of the trampoline (spacetime). Smaller objects, like marbles or light, follow the dips and valleys created by the larger mass, which we interpret as gravity.

Table 20: Science analogies and their origin. (continued)

#	Origin	Metaphor
75	Gemini	Epigenetics is like the annotation layer of a digital book: The underlying text (the DNA sequence) remains fixed and unedited. However, a separate, dynamic layer of highlights, notes, and sticky tabs (epigenetic marks like methylation) dictates which parts of the text are read (transcribed into genes), when they are read, and how loudly they are expressed, effectively altering the story's presentation without changing the words.
76	Gemini	Carbon Dioxide accumulation in the atmosphere is like the slow, silent rise of a bathroom sink's overflow plug: Day after day, we keep the faucet running (emissions), and the water level rises steadily. While the overflow plug (the natural carbon sinks) can handle the initial, gradual rise, its capacity is finite. Once the water crests that plug, the damage (climate change) accelerates rapidly and is irreversible without drastic intervention.
77	Gemini	Editing genes with the CRISPR technology is like the digital "Find and Replace" function in a massive document: The guide RNA is the search query that precisely locates a specific, often minute, sequence of text (the target gene). The Cas9 enzyme is the cursor and editor, capable of cutting out the unwanted text (the faulty gene) and inserting new, corrected text (a functional gene) with unparalleled speed and precision.
78	Gemini	Quantum superposition is like the musical chord played by a single, vibrating violin string: The string is simultaneously holding the potential for every single harmonic and overtone in its possible range of sounds. It is only when an ear (the observer) listens (takes a measurement) that the string collapses from all possibilities into a single, definitive note (a single eigenstate).
79	Gemini	Protein folding is like the deployment of a pre-packed, multi-sectional emergency bridge: The unfolded protein is a long, linear series of linked parts (amino acids). As it emerges, it rapidly and spontaneously snaps into a highly specific 3D structure (the bridge's functional form) guided by chemical interactions, because only that final shape will allow it to cross the gap and perform its specific job.
80	Gemini	The event horizon of a black hole in space is like the boundary where a flowing river's speed exceeds a boat's maximum speed: Before this point, a boat (light or matter) can choose to turn around and escape. Once the river's current (spacetime) is flowing faster than the boat's engine can counteract (the speed of light, c), the boat is inevitably pulled toward the waterfall (the singularity) and can never return, regardless of its efforts.
81	Mistral	Learning in a neural network is like sculptors refining a block of marble: Just as sculptors chip away at marble to reveal a statue, neural networks refine their connections through learning, gradually revealing an optimized model.
82	Mistral	Evolutionary adaptation is like a library reorganizing its books based on reader preferences: Similar to how a library might rearrange its collection to better suit the interests and needs of its patrons, species adapt over time to better fit their environment.
83	Mistral	The Higgs field and particle mass are like a cosmic soup that particles swim through: Just as objects moving through a thick soup encounter resistance, particles gain mass as they interact with the Higgs field.
84	Mistral	Black holes are like cosmic vacuum cleaners: Similar to how a vacuum cleaner pulls in debris, black holes draw in matter and light with their intense gravitational pull.
85	Mistral	A superconductor is like a frictionless ice rink for electrons: Just as skaters glide effortlessly on ice, electrons move through a superconductor without resistance, allowing for perfect conductivity.
86	Mistral	Quantum Entanglement of photon pairs is like a pair of magical walkie-talkies: Just as messages sent through these walkie-talkies are instantly heard by the other, changes in one entangled photon are instantly reflected in its partner, no matter the distance.
87	Mistral	Metaphors in science communication are like bridges made of light: Just as these ethereal bridges connect distant lands, metaphors connect complex scientific ideas to everyday experiences, making them accessible.

Table 21: Science analogies and their origin. (continued)

#	Origin	Metaphor
88	Mistral	The vagus nerve is like the body's internal fiber-optic network: Similar to how fiber-optic cables transmit vast amounts of data quickly, the vagus nerve transmits signals throughout the body, influencing various bodily functions.
89	Mistral	Your body's cells are like tiny biological factories: Just as factories produce and manage goods, cells produce energy and materials, managing the body's needs and functions.
90	Mistral	The brain's default mode network (DMN) is like the brain's screensaver mode: Similar to how a screensaver activates when a computer is idle, the DMN activates when the brain is at rest, facilitating daydreaming and self-reflection.
91	Mistral	Telomeres on chromosomes are like the protective plastic tips on shoelaces: They prevent the lace (DNA) from fraying (degrading) over time, ensuring the lace (chromosome) remains intact and functional for as long as possible.
92	Mistral	Dysplastic cells are like rebellious soldiers in an army: Instead of following orders and maintaining formation (normal cell function), they act unpredictably, disrupting the harmony and potentially causing chaos (tissue dysfunction or cancer).
93	Mistral	Entropy in thermodynamics is like a deck of cards being shuffled repeatedly: No matter how many times you shuffle (energy transfer), the cards (particles) become increasingly disordered, and it's impossible to return them to their original, perfectly ordered state without external effort.
94	Mistral	Spacetime curvature is like a trampoline with a bowling ball placed on it: The ball (a massive object like a planet) creates a depression (curvature) in the trampoline (spacetime), causing smaller objects (like marbles or satellites) to roll or orbit around it.
95	Mistral	Epigenetics is like a library's system of bookmarks and sticky notes: While the books (genes) themselves remain unchanged, the notes (epigenetic markers) determine which books are read (expressed) and when, allowing the library (organism) to adapt to different needs without altering the original texts.
96	Mistral	Carbon Dioxide accumulation in the atmosphere is like adding extra blankets to a bed: Each blanket (CO2 molecule) traps more heat (infrared radiation), making the bed (Earth) warmer over time, but removing them (reducing emissions) becomes harder as they pile up.
97	Mistral	Editing genes with CRISPR technology is like using a word processor's 'find and replace' function: You can locate a specific word (gene sequence) in a document (DNA) and replace it with another, but you must be precise to avoid unintended changes elsewhere in the text.
98	Mistral	Quantum superposition is like a spinning coin: While it's in the air (unobserved), the coin exists in a state of both heads and tails simultaneously, but as soon as it lands (is observed), it 'chooses' one state, collapsing the superposition into a definite outcome.
99	Mistral	Protein folding is like origami: A single sheet of paper (amino acid chain) follows a precise set of folds (interactions) to create a specific 3D shape (protein structure), and even a small misfold can turn the intended crane (functional protein) into a useless or harmful wad of paper.
100	Mistral	The event horizon of a black hole is like a one-way airport departure gate: Once you pass through it (cross the event horizon), there's no turning back (escaping the black hole), and all paths lead inexorably toward the destination (singularity), no matter how fast or in what direction you move.

Creation and Validation of a Monolingual Spanish NLI Dataset for Metaphor Interpretation via Model-in-the-Loop

Alec Sánchez-Montero[1], Gemma Bel-Enguix[2], Sergio-Luis Ojeda-Trueba[2]
[1]Universitat Pompeu Fabra, [2]Universidad Nacional Autónoma de México
alecmisael.sanchez@upf.edu, {gbele, sojedat}@iingen.unam.mx

Abstract

Large Language Models (LLMs) can easily generate fluent text, but assessing whether they truly understand metaphors requires moving beyond English-centric datasets and binary token classification tasks. To test if current state-of-the-art models perform genuine structural alignment and analogical reasoning rather than just echoing statistical token co-occurrence, we introduce a new monolingual Spanish Natural Language Inference (NLI) dataset specifically built for metaphor interpretation. Using a Model-in-the-Loop approach, we reconstruct the literal truth conditions of metaphors sourced from science texts. Before human experts curated the data, we performed an ablation study—evaluated via BERTScore and Cross-Entropy—to test whether explicit symbolic scaffolding improves analogical reasoning. While automated evaluations suggested that forcing models to follow explicit metaphorical rules diminished their fluency and increased text surprisal, human evaluation revealed the opposite: this explicit guidance produced far more accurate and strictly literal outputs. This reveals a limitation in how we evaluate NLU: automated metrics consistently penalize the cognitive 'heavy lifting' required to resolve a metaphor, simply because they are built to reward surface-level statistical fluency. By releasing this resource, we aim to shift the focus from surface-level generation to real cognitive alignment and metaphorical understanding in Spanish NLU.

Keywords: metaphor, natural language inference, automatic metahor interpretation, language resources

1. Introduction

Figurative language expressions, and metaphors in particular, remain a persistent challenge for Natural Language Understanding (NLU). Within this subject, the focus in Natural Language Processing (NLP) has traditionally been on metaphor detection, which has been understood as a classification problem and a sequence labeling task (Rai and Chakraverty, 2021). However, in recent years, literature has addressed metaphor interpretation as a key task for measuring true 'understanding' by Large Language Models (LLMs), arguing that comprehending a metaphor requires going beyond mere lexical detection or contrasting meanings, as proposed by dominant annotation protocols such as MIP/MIPVU (Pragglejaz, 2007; Steen et al., 2010), to involve structural alignment and analogical reasoning (Bowdle and Gentner, 2005). In other words, the focus has shifted from identifying which tokens are used metaphorically to ensuring that models can grasp the underlying semantic conditions that validate a figurative expression (i.e., as opposed to mimicking statistical patterns based on token co-occurrence).

In this context, the Natural Language Inference (NLI) framework has emerged as the benchmark method for assessing metaphorical competence in LLMs. Following established protocols such as FLUTE (Chakrabarty et al., 2022), this task is typically framed as determining whether a literal context (Premise) entails or contradicts a metaphorical expression (Hypothesis) (Stowe et al., 2022; Tong et al., 2024; Sengupta et al., 2025). This approach is a step up from simple statistics or lexical bias; rather than simply paraphrasing a metaphor, the model must demonstrate that it can discern the literal state of affairs that makes a metaphor semantically valid. Reliable metaphor understanding by LLMs is vital for downstream tasks such as machine translation and summarization (Rana et al., 2025).

Nevertheless, advances in this field are constrained by two major challenges. First, most high-quality NLI tasks and datasets (such as FLUTE or MUNCH) focus on the English language. In contrast, the scarcity of comparable resources in languages such as Spanish makes it difficult to evaluate LLMs in specific linguistic and cultural contexts, thereby limiting analysis to imperfect transfer phenomena or reliance on multilingual models (Sanchez-Bayona and Agerri, 2025). Secondly, creating NLI datasets for metaphor understanding involves a complex and time-consuming annotation task.

This paper addresses the scarcity of resources by introducing a monolingual Spanish NLI dataset specifically designed for metaphor interpretation. We adopt a Model-in-the-Loop (MITL) methodology for efficient data annotation: LLMs generate candidate premise-hypothesis pairs, which are then curated by human experts. Furthermore, we conduct an ablation study to probe whether explicit metaphorical guidance triggers better reasoning capabilities in LLMs compared to pure statistical

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 77–87
11 May 2026.

inference, contributing to the discussion on emergent analogical reasoning in small-data scenarios. This work represents the foundational phase of a broader research agenda.

The remainder of this paper is organized as follows: Section 2 reviews the shift from metaphor identification to interpretation within the NLI framework. Section 3 details the MITL methodology we used to annotate our dataset, along with the experimental design and the ablation study regarding explicit metaphorical guidance. Section 4 presents the results and analysis derived from human evaluation, followed by a discussion on the implications for analogical reasoning. Finally, Section 5 summarizes our contributions and future directions.

2. Related Work

2.1. The Concept of Metaphor

Traditionally, linguistic research in NLP has approached metaphor primarily through either Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980) or MIP/MIPVU methodologies (Pragglejaz, 2007; Steen et al., 2010). CMT posits metaphors as fixed, static mappings between a source and a target domain (e.g., ARGUMENT IS WAR), while MIP/MIPVU focuses on a lexical procedure to identify these mappings by contrasting "basic" dictionary meanings with contextual usage. While these frameworks have been instrumental for metaphor detection and corpora annotation, they treat metaphoricity as a binary property of text, largely ignoring the real-time cognitive and communicative mechanisms required for metaphoric competence.

However, when evaluating LLMs, the question is not merely whether a metaphor exists, but *how* it is processed and represented by the system. Unlike the static view of CMT or the binary token-based framework of MIP/MIPVU, the Career of Metaphor (CoM) theory suggests that metaphor is a dynamic process depending on how expressions become conventionalized (Bowdle and Gentner, 2005). According to this framework, metaphor processing shifts from structural alignment (an active, computationally expensive comparison or analogy) in novel metaphors to categorization (retrieval of stored associations of lexical items) as they become conventional.

This distinction is particularly relevant for the Reasoning vs. Statistics debate in NLP (Webb et al., 2023). While LLMs demonstrate prowess in completing figurative patterns, it remains unclear whether this capability stems from genuine structural generalization—akin to the alignment process—or merely from exploiting higher-order statistical correlations found in pre-training data, effectively treating all metaphors as conventionalized categories. If models rely exclusively on distributional priors, they may fail to interpret metaphors that require active mapping between distant domains. Therefore, adopting the CoM framework enables us to probe whether LLMs are simulating reasoning or merely retrieving frequent patterns.

2.2. Literature Review

Recent work has highlighted significant gaps in how NLP systems model figurative language. A fundamental issue lies in the definition of the task itself. As noted by Fuoli et al. (2025), metaphoricity is not a binary label (0/1) but a "radial category with more or less prototypical examples". Consequently, while recent studies evaluating LLMs report high F1 scores in binary identification via fine-tuning, these models operate as "black boxes" that replicate labels without modeling the underlying process. Zero-shot prompting performance remains inconsistent, suggesting that models lack a robust internal representation of metaphoricity (Fuoli et al., 2025).

While early benchmarks like GLUE (Wang et al., 2018) treated NLI as a broad, sentence-level semantic task, Chakrabarty et al. (2022) introduced FLUTE as a framework for figurative language processing, which validated that a model understands a metaphor (Hypothesis) only if it can reconstruct the literal situation (Premise) that sustains its truth conditions. However, recent findings have questioned LLMs' ability to perform this mapping genuinely. Although LLMs perform above chance in Winograd-style tasks, they rely on distributional surface cues rather than deep semantic interpretation, according to Liu et al. (2022).

On the other hand, Chen and Mao (2023) identified that even specialized architectures, such as MelBERT (Choi et al., 2021), exhibit "analogical blindness," which means they can effectively identify the Target domain (the topic) but fail to retrieve the Source domain (the image) of a metaphor, indicating that mere detection does not imply structural mapping or conceptual grounding. In the context of Spanish, Puraivan et al. (2024) observe a similar trend: while GPT-4 excels at disambiguating polysemous verbs, its "interpretations" often collapse into static dictionary definitions—a form of categorization—rather than context-aware conceptual mappings.

Therefore, the core debate is whether LLMs are capable of transcending statistical association through explicit scaffolding. Recently, Sengupta et al. (2025) analyzed how the explicit identification of source and target domains acts as a cognitive "boost" for improving NLI performance in few-shot settings. This dependence on external guidance contrasts with psycholinguistic findings in humans

by Ahrens et al. (2024), who found that explicit metaphor signaling does not necessarily aid in the comprehension of novel expressions.

This behavioral divergence forces us to confront a fundamental hypothesis about the inner workings of current language models. Unlike embodied human cognition for metaphor processing, LLMs may require explicit symbolic markers to emulate the structural alignment process and avoid the pitfalls of statistical categorization or conventionality biases, as observed in Italian LLMs by Mazzoli et al. (2025). While Webb et al. (2023) argue that analogical reasoning has emerged as a zero-shot capability in large-scale models, if LLMs truly possessed the flexible structural alignment seen in human cognition, they would not require the explicit symbolic scaffolding observed in recent NLI tasks. Therefore, the reported "analogical prowess" of these models may be less an emergent cognitive faculty and more a reflection of their ability to navigate higher-order statistical correlations within the lexicon.

3. Resource Creation

The development of this resource stems from the need to move beyond surface-level label detection toward the resolution of underlying semantics in Spanish. To this end, we adapted the NLI task structure from FLUTE (Chakrabarty et al., 2022; Sengupta et al., 2025) to a curated set of metaphorical expressions. The final annotated dataset is publicly available on a GitHub repository.

3.1. Sample Selection and Curation Process

The source data consists of Spanish tweets focused on Popular Communication of Science (PCS), originally annotated by Sánchez-Montero et al. (2025). To ensure experimental validity and control, we applied a strict filtering criterion: we selected only the 200 instances that achieved "perfect human agreement" for the overall metaphor category (i.e., a soft-label score of 1.0). We hypothesized that using these prototypical examples would isolate the model's reasoning capacity by eliminating the semantic noise inherent in borderline cases of metaphoricity. This sample includes various scientific and colloquial metaphors used in PCS, such as personifications (e.g., telescopes making new scientific discoveries), direct comparisons (e.g., presenting evolution as a tree), and lexical metaphors (e.g., black holes or digital threads).

For each selected instance, we identified the specific tokens that were previously annotated as metaphorical. We then manually authored a brief natural language explanation (NLE) for each metaphor, adhering to the classification schema used in the original dataset (e.g., Direct, Indirect, or Personification). These NLEs constitute the ground-truth metaphorical knowledge, which is a critical component of our experimental design, as it provides the explicit symbolic guidance required for the ablation study, where the presence or absence of this expert-level information serves as the primary independent variable.

3.2. NLI Annotation via Model-in-the-Loop

Following the paradigm established by Chakrabarty et al. (2022), the NLI resource we introduce is structured as a reconstruction of truth conditions. For every metaphorical Hypothesis (H), we generated two types of literal Premises (P):

- **Entailment (E):** A literal description of the real-world situation that validates the figurative expression.
- **Contradiction (C):** A literal scenario that renders the analogical mapping proposed in the hypothesis false or impossible.

We utilized a *model-in-the-loop* approach to generate these premises, employing `gpt-4.1-mini` (OpenAI et al., 2024), `Qwen2.5-7B-Instruct` (Quen: An Yang et al., 2025), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), and `gemma-3-12b-it` (Team et al., 2025) as the generative engines. With this selection, we aimed to evaluate whether metaphorical reasoning scales across different architectures and paradigms. While GPT-4.1 serves as a highly aligned proprietary baseline, the inclusion of robust open-weight models (Qwen, Mistral, and Gemma) allows us to investigate whether the capacity for structural alignment in Spanish is a generalized emergent capability or is heavily dependent on proprietary instruction-tuning pipelines.

The premise generation followed two distinct configurations for our ablation study:

- **Baseline Configuration:** The model received only the metaphorical Hypothesis (H) and was tasked with generating E and C without any external guidance.
- **Scaffolded Configuration:** The model received H along with the NLE regarding the identified metaphors. This setup forces the model to integrate external symbolic knowledge before performing the inference task.

We implemented a few-shot prompting strategy, providing two prototypical examples of (H, E, C) triplets. We included these examples to anchor the model's output to a literal register, explicitly forbidding the use of figurative language in the premises.

Drawing on the experimental framework of Sengupta et al. (2025), who utilized a temperature of 0.8 for metaphorical NLI tasks, we opted for a more constrained temperature of 0.7. This setting facilitates sufficient linguistic variation for Spanish paraphrasing while maintaining the logical precision required for the premises. During our initial pilot tests, higher temperatures led to a degradation in literalness, causing the models to hallucinate secondary metaphorical associations rather than produce strictly factual premises. Each prompt was executed in an independent session to eliminate the risk of context-window leakage or intra-experimental bias.[1]

3.3. Automatic Quality Estimation

Given the labor-intensive nature of the human validation step in the MITL pipeline, manually reviewing the outputs from every model and configuration would be highly impractical. Therefore, to focus our annotators' efforts and identify the most reliable option for the final curation, we implemented a preliminary automatic evaluation phase. We assessed the generated entailments (E) using two complementary metrics:

- **Semantic Coherence (BERTScore):** We used BERTScore (Zhang et al., 2020) with the multilingual `xlm-roberta` checkpoint (Conneau et al., 2020) to measure how well the generated entailment (E) preserved the underlying meaning of the original metaphorical text (H). The metric ranges from 0 to 1, with higher values reflecting a stronger semantic overlap between E and H. Additionally, we implemented a strict penalty for "language drift"; any premise generated in English received a significant deduction to prevent high scores from being assigned to non-Spanish outputs.
- **Conditioned Fluency and Information Loss (Cross-Entropy):** We calculated the average Cross-Entropy (CE) loss to evaluate the predictability and linguistic quality of the literal E conditioned on the original metaphorical H. Rather than evaluating the generated text in isolation, we measured the surprisal of E given H. To avoid self-preference bias, we employed two independent causal models—`Llama-3.1-8B` (Grattafiori et al., 2024) and `Mistral-7B` (Jiang et al., 2023)—as external judges. Lower CE scores represent lower surprisal when transitioning from the metaphorical source to the literal target, indicating a more predictable and fluent reconstruction of truth conditions and a minimal loss in Spanish.

The results of this automatic evaluation across both the baseline and scaffolded configurations are detailed in Table 1.

As detailed in Table 1, performance varied depending on the models' ability to maintain the target language. Mistral-7B, for instance, exhibited language drift in over 20% of its outputs. Due to our strict penalty for code-switching, its overall BERTScore fell to 0.7589, well below the rest of the group. At the other end of the spectrum, the GPT baseline delivered the most stable results; it achieved the highest semantic overlap (BERTScore = 0.847) and produced the most predictable Spanish phrasing, reflecting the lowest CE values from both judges (1.2842 and 1.2578).

A critical takeaway from the ablation setup is the behavior of the scaffolded configuration, which constitutes an unexpected pattern regarding explicit guidance. Except for Mistral-7B, providing the explicit NLE led to higher cross-entropy and lower BERTScores across all models. This might indicate that processing explicit metaphorical constraints disrupts their fluency and leads to more literal metaphor paraphrases that could be perceived as less natural. Consequently, we discarded the 'noisier' outputs and exclusively forwarded the GPT-generated premises to the human curation phase, which meant that our expert annotators could focus entirely on evaluating complex logical and analogical errors instead of filtering out basic translation or grammatical failures.

4. Manual Validation

4.1. Overall Assessment

Although automated metrics like BERTScore and Cross-Entropy effectively filter out low-quality outputs, they ultimately act as proxies for surface-level fluency and lexical overlap. These scores cannot definitively prove whether a model has genuinely 'understood' a metaphor. Therefore, to assess the top performer's metaphorical competence beyond statistical measurements, we conducted a manual validation via expert annotation. Based on the automated quality estimations, we isolated the outputs from our top-performing model (GPT-4.1) to build our core NLI dataset.

During this phase, we annotated the generations from both the baseline and the scaffolded configurations. We anticipated that a side-by-side comparison of these setups would reveal how injecting external symbolic knowledge (via the NLEs) shifts the model's underlying interpretative process. Still, since expert manual evaluation of semantics and

[1] The complete translated text of the prompts, including the few-shot examples used for both configurations, is detailed in the Appendix.

Model / Configuration	BERTScore (↑)	CE (Llama 3.1) (↓)	CE (Mistral 7B) (↓)
GPT Baseline	**0.8470**	**1.2842**	**1.2578**
GPT Scaffolded	0.8421	1.3724	1.3178
Qwen Baseline	0.8449	1.4192	1.3745
Qwen Scaffolded	0.8400	1.5986	1.5333
Mistral Baseline	0.7589	1.4852	1.2911
Mistral Scaffolded	0.7777	1.5630	1.3684
Gemma Baseline	0.8332	1.4928	1.4488
Gemma Scaffolded	0.8305	1.6026	1.5286

Table 1: Automatic Evaluation of E premises in relation to H: BERTScore and Cross-Entropy

pragmatics is incredibly time-consuming, we narrowed the scope of this initial validation phase. We focused our qualitative analysis on the 200 items from the dataset, restricting our review entirely to the Entailment (E) premises.

We opted to leave the Contradiction (C) premises for future work due to the limitations in how current models handle metaphorical negation. In our preliminary experiments, we observed that the model couldn't reliably negate just the metaphors in H. Instead, it usually ended up bluntly negating the entire situation. Validating the C premises requires checking whether a model can actively negate a metaphorical mapping without breaking basic logic—a much more demanding task that could have derailed the manual validation, given the quality of these outputs. Therefore, isolating the entailments ensured our annotators could zero in on one core question: did the model accurately and literally reconstruct the truth conditions that validate the metaphor?

We graded the generated entailments using a 3-point scale (0, 0.5, and 1). We deliberately avoided a simple binary system because LLM outputs are highly nuanced; a model might successfully unpack one metaphor but stumble on another within the same sentence. Our annotation criteria were defined as follows:

- **Score 1 (Fully Correct):** To get a perfect score, the generated text had to completely strip away all figurative language (i.e., it had to be 100% literal) while accurately and comprehensively explaining the underlying meaning of the metaphor (or all the metaphors, if multiple were present) from the original hypothesis.

- **Score 0.5 (Partially Correct):** Assigned when the model grasped the core meaning but failed to execute the task flawlessly. This happened mostly in two scenarios: either the model explained the metaphor but accidentally snuck in a new figurative expression, or it successfully resolved one metaphor but missed others present in the same context.

- **Score 0 (Incorrect):** Reserved for clear failures in metaphorical reasoning. We assigned a zero if the model hallucinated facts, logically contradicted the source text, or just "parroted" the original figurative words instead of translating them into a literal reality.

The distribution of human-annotated scores across the two experimental configurations is detailed in Table 2.

The manual validation seems to contradict our preliminary automated metrics. On paper, the unguided baseline configuration appeared superior: it generated text with higher statistical fluency (lower cross-entropy) and better lexical overlap (higher BERTScore). Yet, when evaluated for actual semantic validity, it failed to capture the literal truth conditions in roughly 24.5% of its generations. The scaffolded configuration, on the other hand, despite scoring lower on automated metrics, proved vastly superior in human evaluations: fully correct entailments climbed from 67.0% to 88.5%, and complete conceptual errors were reduced by nearly two-thirds, dropping from 24.5% to 8.5%.

This clash between statistical metrics and human judgment reveals a troubling blind spot in NLU evaluation for figurative language. The explicit scaffolding appears to 'force' the model to engage in active structural mapping rather than relying on highly probable word associations. Doing this cognitive "heavy lifting" naturally increases text surprisal and lowers lexical overlap metrics. Nonetheless, as the human scores demonstrate, accepting this higher surprisal is precisely what is required to generate factually accurate and literal premises and, ultimately, to verify whether current models can actually understand and interpret real metaphorical language.

4.2. Qualitative Analysis

When we manually reviewed the outputs where the zero-shot baseline failed but the scaffolded model succeeded, two distinct behavioral patterns emerged. Without explicit guidance on how to handle figurative language, the baseline GPT model

Configuration	Score 1	Score 0.5	Score 0
GPT Baseline	134 (67%)	17 (8.5%)	49 (24.5%)
GPT Scaffolded	**177 (88.5%)**	6 (3%)	**17 (8.5%)**

Table 2: Human evaluation results for generated entailments ($N = 200$), comparing the zero-shot baseline against the explicitly guided scaffolded configuration.

essentially fell into two traps: it either ignored the metaphor completely (contextual distraction) or it tried to explain it but accidentally used another metaphor (analogical blindness).

1. Contextual Distraction and Omission

The first common error was simple omission. Faced with a metaphorical expression it did not quite know how to map, the unguided model often took the easy way out: it ignored the figurative phrase entirely or just summarized the literal context surrounding it. Since the actual truth conditions of the metaphor were lost, these outputs received 0.

Consider this example of personification from the dataset:

> **Original Text (ES):** *"...la hazaña fue posible luego de que el orbitador de la misión Tianwen-1 analizara el suelo marciano durante tres meses."*
> **Original Text (EN):** *"...the achievement was made possible after the Tianwen-1 mission orbiter analyzed the Martian soil for three months."*
> **Metaphor:** Personification (attributing the cognitive action of 'analyzing' to a machine).
>
> **Baseline Generation (Score 0) (ES):** *"China envió una nave espacial que aterrizó con éxito en Marte y desplegó un rover llamado Zhurong para explorar la superficie..."*
> **Baseline Generation (Score 0) (EN):** *"China sent a spacecraft that successfully landed on Mars and deployed a rover named Zhurong to explore the surface..."*
>
> **Scaffolded Generation (Score 1) (ES):** *"Un satélite en órbita alrededor de Marte recopiló y transmitió datos sobre la composición y características del suelo marciano..."*
> **Scaffolded Generation (Score 1) (EN):** *"A satellite orbiting Mars collected and transmitted data on the composition and characteristics of Martian soil..."*

In this instance, we observe a clear case of semantic displacement. Although the baseline model recognizes the general topic (China's Mars mission), it fails to resolve the specific metaphorical action of the orbiter. Instead of translating the verb *analizara* (analyzed) into its physical equivalent, the model defaults to generating a high-probability summary of the overall event. In contrast, the scaffolded configuration uses the explicit NLE to anchor its attention: by being told that "analyzing" in this context refers to a mechanical process, the model is forced to identify that an orbiter "analyzes" soil by *collecting and transmitting data*.

This omission behavior was not limited to scientific concepts; it also affected digital metaphors. In another instance involving a Twitter thread (*"Te comparto un hilo con información..."* = Here's a thread with some information...), the baseline completely omitted the concept of the thread, generating a premise about showing an image instead. As one of our evaluators explicitly flagged in the annotation logs: *"[The model] completely ignores 'thread'."* The scaffolded version, conversely, successfully mapped the metaphor to its literal reality: *"Se publicaron varios mensajes relacionados entre sí..."* (Several interconnected messages were published).

When faced with sentences containing multiple layered metaphors, the unguided baseline often suffered from "metaphorical fatigue." It successfully unpacked the first figurative expression but completely ignored the subsequent ones. Consider this example regarding the neurochemistry of love:

> **Original Text (ES):** *"...tu cerebro produce oxitocina, dopamina y serotonina, un verdadero cóctel de felicidad. Además, se activa constantemente el 'circuito de recompensa'..."*
> **Original Text (EN):** *"...your brain produces oxytocin, dopamine, and serotonin, a real cocktail of happiness. In addition, the 'reward circuit' is constantly activated..."*
> **Metaphors:** Metaphorical comparison (*cóctel*), personification (*produce*), and lexicalized metaphor (*circuito de recompensa*).
>
> **Baseline Generation (Score 0.5) (ES):** *"...el cerebro libera oxitocina... y se activa el **circuito de recompensa** asociado con sensaciones placenteras."*
> **Baseline Generation (Score 0.5) (EN):** *"...the brain releases oxytocin... and activates the **reward circuit** associated with pleasurable sensations."*
>
> **Scaffolded Generation (Score 1) (ES):** *"...el cerebro libera simultáneamente oxitocina... además, se activa una **red neuronal** asociada a la recompensa."*
> **Scaffolded Generation (Score 1) (EN):**

> *"...The brain simultaneously releases oxytocin... In addition, a **neural network** associated with reward is activated."*

As our annotator log points out: *"it completely failed to find the second metaphor."* While the baseline correctly translates the "cocktail" into a literal chemical release, it becomes distracted and leaves the lexicalized "reward circuit" intact.

2. Analogical Blindness

The second error pattern identified is a form of analogical blindness (Chen and Mao, 2023), where the model "interprets" a metaphor by replacing it with another equally figurative expression.

Consider this example of personification from the dataset:

> **Original Text (ES):** *"El telescopio espacial Spitzer ha descubierto 120 000 nuevas estrellas recién nacidas..."*
> **Original Text (EN):** *"The Spitzer Space Telescope has discovered 120,000 new newborn stars..."*
> **Metaphors:** Personification (attributing the action of 'discovering' to a machine and the property of being 'newborn' to stars).
>
> **Baseline Generation (Score 0.5) (ES):** *"El telescopio espacial Spitzer detectó radiación infrarroja proveniente de nubes de polvo en el plano galáctico y, a partir de esos datos, identificó 120 000 estrellas **jóvenes** recién formadas."*
> **Baseline Generation (Score 0.5) (EN):** *"The Spitzer Space Telescope detected infrared radiation coming from dust clouds in the galactic plane and, based on this data, identified 120,000 newly formed **young** stars."*
>
> **Scaffolded Generation (Score 1) (ES):** *"Un telescopio espacial detectó señales de radiación infrarroja provenientes del polvo en las nubes interestelares, lo que permitió a los científicos identificar 120 000 estrellas que se formaron recientemente en el plano de la galaxia."*
> **Scaffolded Generation (Score 1) (EN):** *"A space telescope detected infrared radiation signals coming from dust in interstellar clouds, allowing scientists to identify 120,000 stars that formed recently in the plane of the galaxy."*

In this paradigmatic case, the baseline (0.5) replaced the term 'newborn' with 'young' stars. Although this is a common astronomical description, from a strictly linguistic perspective, "young" is still a metaphor of personification projected onto an inanimate object. As the evaluators point out in the logs: "it replaces newborn with young, another metaphor." In contrast, the scaffolded model (1.0) managed to break the analogical cycle by describing the physical process as "stars that were recently formed."

Nonetheless, our analysis uncovered a blurred line where sometimes absolute literalness seems to border on the impossible. In cases of highly lexicalized or abstract metaphors, such as "black holes" or the very concept of technological 'tools,' even the scaffolded model struggled. A critical finding in manual validation was the persistence of personification in scientific instruments, attributing them the capabilities of "analyzing," "discovering," and "observing." Technological tools, such as telescopes, are essentially extensions of the human body; we attribute human characteristics to them because they allow us to perform human actions on a non-human scale. In these cases, the only way to achieve literalness was to shift the agency from the object to the subject, as can be seen in the following example:

> **Original Text (ES):** *"El telescopio Webb analizó la estrella Fomalhaut..."*
> **Original Text (EN):** *"The Webb telescope analyzed the star Fomalhaut..."*
> **Metaphors:** Personification (attributing the action of 'analyzing' to a machine).
>
> **Scaffolded Generation (Score 1) (ES):** *"Los científicos utilizaron el telescopio Webb para observar la estrella Fomalhaut..."*
> **Scaffolded Generation (Score 1) (EN):** *"Scientists used the Webb telescope to observe the star Fomalhaut..."*

Sometimes, as can be seen in the example, interpretive success did not consist of finding a literal synonym for a particular verb, but rather of reconstructing the entire scene. The symbolic guidance via NLE seems to be the only mechanism that forces the model to recognize this hierarchy of human agency, preventing the model from merely "parroting" the linguistic convention of personifying technology.

Finally, there were a few remaining cases where the model simply failed to decipher the text-context relationship. Without symbolic scaffolding, models tend to "hallucinate" explanations based on the statistical frequency of adjacent words, losing sight of the truth condition that underpins the analogy. Despite their fluency, LLMs often get stuck on the surface of language, needing a symbolic "push" to understand metaphorical language.

4.3. Discussion

We have observed that forcing the model to integrate explicit knowledge of NLEs increases the cross-entropy of the resulting text and reduces its lexical overlap (BERTScore). For traditional evaluation systems, this increase in statistical surprise is a

symptom of degradation; however, human evaluation shows that it is precisely the cognitive cost necessary to achieve semantic accuracy when dealing with metaphor understanding.

This empirical tension supports Liu et al. (2022), as the apparent semantic competence of LLMs is often based on distributional clues rather than deep context understanding or genuine analogical reasoning. When the model operates without scaffolding, its natural inertia is to remain within the statistical comfort zone of figurative language. Replacing the concept of "newborn stars" with "young stars" is not a step toward literalism, but rather a lateral shift within the same semantic domain, which may indicate that it did not completely understand the meaning of the metaphor. In terms of Bowdle and Gentner (2005), this analogical blindness might be explained because models treat metaphors as static categories and retrieve frequent synonyms rather than performing structural alignment between different domains. In other words, metaphorical knowledge might be coded as categorization rather than instances of analogy.

Symbolic scaffolding acts precisely as the mechanism to break this inertia. Forcing the model to transition from statistical categorization to analogical inference also reveals an epistemological boundary in the evaluation task itself. Reaching the "zero degree" of literalness often clashes with how we inherently conceptualize conceptual domains such as science and technology. Historically, we have endowed our scientific instruments with human agency because they function as physical extensions of our own bodies. Consequently, in natural language, telescopes "look," rovers "analyze," and probes "discover." This assimilation of metaphor as a simple static category not only reflects a computational limitation, but also raises the question about the importance of embodied cognition in the architecture of NLP systems and, by extension, in the pre-training data of models.

For the unguided model, the metaphoricity of some expressions remains invisible due to its overwhelming frequency in the pre-training data. The metaphor is so ubiquitous that it masquerades as literal truth. Providing explicit scaffolding, however, forces a pragmatic shift. Rather than just swapping vocabulary, in some cases, the guided model actively rewrites the text to hand agency back to the human researchers (e.g., *scientists used the telescope to observe*...). Interpreting a metaphor in an NLI context goes far beyond simple intralinguistic translation or synonym hunting, as it requires the model to step outside the text and actively reconstruct the real-world hierarchy of who acts and what is acted upon.

5. Conclusion and Future Work

In this work, we have presented the first monolingual Natural Language Inference (NLI) dataset in Spanish, dedicated specifically to metaphor interpretation. Through a Model-in-the-Loop methodology, we have demonstrated that it is possible to generate accurate reconstructions of the truth conditions of figurative expressions, provided that an explicit symbolic scaffolding is supplied.

The method shows promising results and suggests a clear direction for future work through the evaluation of LLM behavior in the detection, classification, and inference of metaphor. Several lines of future research are identified. Also, it will be necessary to investigate different prompting strategies and forms of contextual knowledge that could enhance LLMs efficiency. On the other hand, experiments must be scaled to larger and more general corpora. It will also be essential to develop hybrid human-LLM evaluation methods and metrics capable of accurately assessing model performance on this task and rewarding genuine cognitive alignment over mere statistical fluency.

Our immediate next step involves the manual validation of the Contradiction (C) premises, a logical challenge that will test whether models can actively negate an analogical mapping without violating real-world constraints. Expanding the scope of this research, we also plan to conduct an in-depth qualitative comparison across different LLMs to understand how their baseline interpretative strategies differ when handling figurative language. Furthermore, scaling these experiments beyond science communication will be essential. Future work must investigate whether model performance and "analogical blindness" shift when encountering highly conventionalized or lexical metaphors in general-domain Spanish texts.

Acknowledgements

This paper has been supported by PAPIIT Project IG400325, as well as a Postgraduate Scholarship by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) (CVU 1225477).

Ethical considerations and limitations

The primary drawback of this paper is the limited size of the dataset ($N = 200$) on which the experiment and subsequent manual evaluation were conducted. Although the results were extensively and rigorously validated by expert annotators, the limited number of instances may restrict the broader generalizability of our findings.

The principles of the Belmont Report (Belmont, 1978) were followed in the creation of the dataset

and in the data annotation process.

6. Bibliographical References

Kathleen Ahrens, Christian Burgers, and Yin Zhong. 2024. Making the unseen seen: The role of signaling and novelty in rating metaphors. *Journal of Psycholinguistic Research*, 53(3):36.

Informe Belmont. 1978. Principios éticos y directrices para la protección de sujetos humanos de investigación. *Estados Unidos de Norteamérica: Reporte de la Comisión Nacional para la Protección de Sujetos Humanos de Investigación Biomédica y de Comportamiento*.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. Association for Computational Linguistics.

Zi-Yuan Chen and Yining Mao. 2023. MetaMapper: Interpretable metaphor detection.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

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.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et *al.* 2024. The llama 3 herd of models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

George Lakoff and Mark Leonard Johnson. 1980. *Metaphors We Live By*. University of Chicago press.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452. Association for Computational Linguistics.

Simone Mazzoli, Alice Suozzi, and Gianluca Lebani. 2025. Language models and the magic of metaphor: A comparative evaluation with human judgments. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 710–721. CEUR Workshop Proceedings.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and Janko Altenschmidt et *al.* 2024. Gpt-4 technical report.

Pragglejaz. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Eduardo Puraivan, Irene Renau, and Nicolás Riquelme. 2024. Metaphor identification and interpretation in corpora with ChatGPT. *SN Computer Science*, 5(8):976.

Quen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Sunny Rai and Shampa Chakraverty. 2021. A survey on computational metaphor processing. *ACM Computing Surveys*, 53(2):1–37.

Manisha Rana, Rita Chhikara, and Srishti Sharma. 2025. A SURRVEY ON METAPHOR DETECTION AND INTERPRETATION. *International Journal For Multidisciplinary Research*, 7(4):54959.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation.

Meghdut Sengupta, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, Debanjan Ghosh, and Henning Wachsmuth. 2025. Investigating the impact of conceptual metaphors on LLM-based NLI through shapley interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17393–17403, Suzhou, China. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in spanish. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 45–56. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and Thomas Mesnard et *al.* 2025. Gemma 3 technical report.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Appendix

Prompt for Baseline Task (English Translation)

You are a linguist specializing in Spanish semantics and pragmatics. Your task is to generate high-quality Natural Language Inference (NLI) pairs.

Instructions:

You are provided with an input sentence containing a metaphorical expression. This sentence serves as the Hypothesis (H).

Your job is to reconstruct the literal context that would generate that hypothesis. You must generate two sentences that function as literal premises:

1. Entailment (E): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis true (i.e., what actually happened in the physical world for someone to use that metaphor).

2. Contradiction (C): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis false or impossible.

Examples:

Example 1:

Input (H): 'The street was a zoo this morning.'

Output:

Entailment: 'There was deafening noise and people were running around in a chaotic environment.'

Contradiction: 'The urban environment was completely silent and orderly, and everyone was walking along quiet.'

Example 2:

Input (H): 'The candidate tried to shield his public image.' Output:

Entailment: 'The politician took extreme measures to protect his reputation and avoid any criticism or attacks from the press.'

Contradiction: ‘The politician openly exposed himself to criticism and shared compromising information about his private life.’

Important:

- The generated premises (E and C) must not contain metaphors. They should be brief and direct explanations.
- Do not repeat the input text.
- Do not repeat or explain the reasoning.

Input:
{text}
Output format:
[ENTAILMENT]
(Write the E sentence here)
[CONTRADICTION]
(Write the C sentence here)

Prompt for Scaffolded Task (English Translation)

You are a linguist specializing in Spanish semantics and pragmatics. Your task is to generate high-quality Natural Language Inference (NLI) pairs based on the key information provided to you.

Instructions:

You are provided with an input sentence containing a figurative or metaphorical expression. This sentence serves as the Hypothesis (H). You are also provided with the metaphors identified in H as key information.

Your job is to reconstruct the literal context that would generate that hypothesis. You must generate two sentences that function as literal premises:

1. Entailment (E): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis true (i.e., what actually happened in the physical world for someone to use that metaphor).
2. Contradiction (C): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis false or impossible.

Examples:

Example 1:

Input (H): ‘The street was a zoo this morning.’

Metaphor: ‘The term “zoo” is used here figuratively.’

Output:

Entailment: ‘There was deafening noise and people were running around in a chaotic environment.’

Contradiction: ‘The urban environment was completely silent, orderly, and everyone was walking around quiet.’

Example 2:

Input (H): ‘The candidate tried to shield his public image.”

Metaphor: ‘The verb “shield” is used here in a figurative sense.’

Output:

Entailment: ‘The politician took extreme measures to protect his reputation and avoid any criticism or attacks from the press.’

Contradiction: ‘The politician openly exposed himself to criticism and shared compromising information about his private life.’

Important:

- The generated premises (E and C) must not contain metaphors. They should be brief and direct explanations.
- Do not repeat the input text.
- Do not repeat or explain the reasoning.

Input:
{text}
Key information: {metaphor}
Output format:
[ENTAILMENT]
(Write the E sentence here)
[CONTRADICTION]
(Write the C sentence here)

A Hybrid Architecture for Metonymy Detection in Marathi

Pratibha Dongare
The English and Foreign Languages University
pratibhaphdlandp22@efluniversity.ac.in

Abstract

Metonymy, often considered as a figurative trope, is a frequently occurring linguistic phenomenon in which an entity is replaced by a semantically related entity. Named entities are commonly used to refer to associated concepts. For instance, in the sentence *India signed a treaty*, the geographical name *India* stands metonymically for the government rather than the physical location. This study develops a hybrid architecture to classify literal and metonymic usages of named entities in Marathi language using small data. The approach integrates Pustejovsky's Generative Lexicon framework with linguistic features, including part-of-speech tags, named entity labels, and lemmas. The model is evaluated on 890 sentences and achieved F1 scores of 66.98% and 71.97% for literal and metonymic instances, respectively. The study highlights the effectiveness of the features in capturing metonymic contexts, though precision remains a target for improvement. Ablation results confirm that the Formal and Constitutive Qualia roles are the most critical components for detecting metonymic shifts, while the Telic role introduces modest noise in the present corpus. This experiment shows the scope for developing hybrid models for learning non-literal language using small data, which could be beneficial for less-explored and low-resource languages.

Keywords: Metonymy, Generative Lexicon, Marathi language

1. Introduction

Named entity recognition models often classify entities as flat and surface level entities. While this helps in classifying NEs, not all entities convey their literal sense. Metonymy, a semantic phenomenon where an entity refers to something associated with it in the real world, occurs frequently in everyday language. Surface level classification overlooks metonymy, which is crucial for various NLP tasks. A location might represent an organization or an event, shifting its semantic role based heavily on the surrounding context. For example, in the sentence *India won the match*, *India* is referring to a sports team rather than a physical location. This is a metonymic instance. The present study examines metonymic senses in NEs using Marathi language data. Marathi is an Indo-Aryan language spoken in the central-western region of India. Although several NLP studies have addressed this language, metonymy detection remains under-explored. The limited availability of annotated datasets for metonymy detection in low-resource languages such as Marathi, necessitates approaches that perform effectively with small data.

This paper evaluates and explores the effectiveness of detecting metonymic shifts in Named Entities using a small dataset. By maintaining a highly controlled annotation environment, we investigate the extent to which a model can accurately learn to distinguish between literal and metonymic readings of NEs. The next section briefly overviews metonymy detection studies. The third section discusses the methodology employed in the present study, while the fourth section presents the experimental results and analysis of the proposed small-data approach. The fifth section discusses the limitations of the present study and finally the paper concludes with potential scope for future research.

2. Related works

Metonymy has been studied in various fields including rhetorics, cognitive linguistics and NLP. In natural language processing, understanding metonymy is critical for tasks such as information extraction and named entity recognition, often framed as a classification problem where a system identifies metonymic usage of a target word within a sentence (Ghosh and Jiang, 2025). Historically, approaches to metonymy resolution have largely depended on manually curated lexical resources, including parsers, taggers, and dictionaries (Gritta et al., 2017). Fass (Fass, 1988; Fass et al., 1997) proposed a method for identifying metonymy based on selectional restrictions, while later research introduced more sophisticated statistical and machine learning techniques. Some researchers have worked on unconventional metonymy (Ghosh and Jiang, 2025) while, some researchers have focused on conventional metonymy detection which accounts for metonymic interpretations of named entities (Poibeau, 2006). Early work in metonymy resolution often focused on selectional preferences, yet this approach frequently overlooked a significant number of metonymic readings (Meguellati et al., 2022). Few scholars have studied metonymy as a classification task (Markert and Nissim, 2002a,b, 2007). Markert & Hahn (Markert and Hahn, 2002) reported 17% of metonymic in-

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 88–92
11 May 2026.

stances in German magazines. Hence, metonymy is prevalent and is a crucial linguistic phenomenon for natural language understanding. (Gritta et al., 2017) introduced a novel dataset, RelocaR, specifically designed to address metonymy in geographical named entities, aiming to improve upon prior annotations. Additionally, they used a novel predicate window approach which used only a small, focused segment of the sentence surrounding the entity's head dependency for classification, thereby enhancing accuracy by minimizing irrelevant input. (Kupelioglu et al., 2016) used named entities, argument structures and wordnet for metonymy resolution. Their approach integrated these diverse linguistic features to capture the nuances of metonymic expressions, demonstrating a significant advancement in automated metonymy detection. The literature suggests that while extensive resources and sophisticated models have been developed for metonymy detection, there remains a critical need for effective strategies in low-resource settings.

3. Methodology

For this study, a dataset of 890 sentences (13807 tokens) was manually annotated for lemmas, part-of-speech, named entities, and senses. Person, Location, Organization, and Miscellaneous tags were used to annotate named entities. The data was sourced from Marathi news articles covering politics, sports, finance, and current affairs domains. Annotation was performed solely by the author, following guidelines from the larger dataset under development. NEs were tagged using BILOU encoding, with each assigned a literal or metonymic sense. The dataset contained approximately 11.9% of NE tokens, of which 54.4% were literal and 45.5% metonymic.

Qualia roles from the Generative Lexicon (Pustejovsky, 1991) were mapped to NEs via a 4-bit binary vector to encode additional potential meanings beyond literal senses. Qualia roles consist of Formal, Constitutive, Telic, and Agentive roles. These roles capture distinct aspects of an entity's semantic structure and help the model search for potential novel meanings a named entity may carry in a given context. Qualia roles were implemented as deterministic, rule-based feature functions rather than learned parameters. For each named entity token, a fixed 4-bit binary vector was computed at preprocessing stage. For example, a LOCATION entity receives Formal=1 (it has a definitional physical boundary), Constitutive=0 (it is not inherently part of a containing structure), Telic=0 (it has no canonical functional purpose), and Agentive=0 (it is not the result of a deliberate causal process). Table 1 summarises the full role-to-bit mapping used in this study. The proposed system employs a BiLSTM-CRF architecture utilizing MahaBERT (Joshi, 2022) contextual embeddings, and was implemented using the PyTorch (Paszke et al., 2019) framework. figure 1 illustrates the system architecture. The annotated data is processed through the feature engineering where linguistic features and embeddings are used. Since Marathi uses a free word order, these features are important cues in detecting metonymy. The BiLSTM layer reads the data forward and backward before assigning labels such as literal or metonymic.

Qualia Role	Bit	Encodes	Example rule for Marathi NEs
Formal	1	What the entity is (type/category)	LOCATION → physical boundary entity
Constitutive	2	What it is made of or part of	ORGANIZATION → comprises people and roles
Telic	3	Its purpose or function	ORGANIZATION → has institutional goal
Agentive	4	Its causal/agentive origin	PERSON → agent of volitional action

Table 1: Qualia role encoding

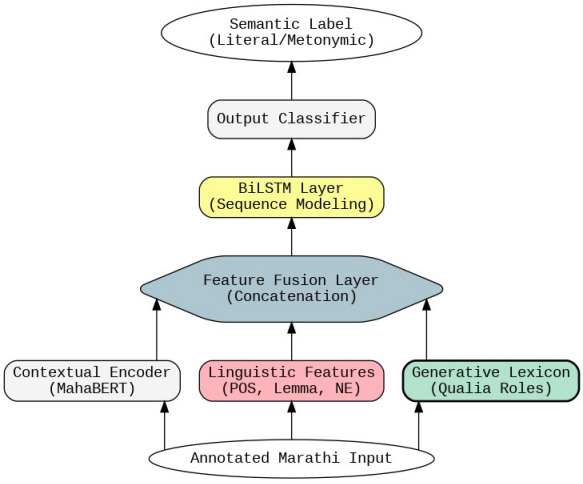


Figure 1: System architecture

4. Results and Discussion

To evaluate the hybrid architecture, we examined token-level predictions. The model achieved a Literal F1 of 66.88 % (Precision: 52.99, Recall: 91.03) and a Metonymic F1 of 71.97% (Precision: 58.50, Recall: 93.48), yielding a Macro F1 of 69.42% across 170 NE instances. Table 2 shows the results. The recall is very high for both classes,

Class	Precision	Recall	F1
Literal	52.99	91.03	66.98
Metonymic	58.50	93.48	71.97
micro avg	55.87	92.35	69.62
macro avg	55.74	92.25	69.42
weighted avg	55.97	92.35	69.68

Table 2: Performance metrics of the model

Features	Precision	Recall	F1
Qualia (all)	59.9	61.2	60.5
Formal	59.9	59.3	59.6
Constitutive	66.7	60.3	63.5
Telic	68.9	70.9	69.9
Agentive	58.8	62.6	60.7
PoS	59.0	63.8	61.8
Lemma	64.1	63.7	63.9
NE	64.9	71.0	67.9

Table 3: Feature wise ablation results.

while precision is comparatively low. This asymmetry indicates that the model is broadly inclusive. It successfully recovered the vast majority of metonymic and literal instances, but overpredicted both labels. Of the NE instances evaluated, 86 were correctly identified as metonymic (true positives) and 71 as literal (true negatives), with only 3 false positives and 5 false negatives at the NE level. Notably, 114 non-NE tokens were predicted as either metonymic or literal by the model contributing to the low overall precision. Figure 2 shows the performance of the model.

4.1. Ablation Study

To assess the individual contribution of each feature, a systematic ablation study was conducted. Each feature was removed independently and the model was retrained; Macro F1 was compared against the full model baseline (Macro F1 = 61.4). Table 3 summarises the results. Several patterns can be seen from this test. Among the Qualia roles, the Formal role had the largest detrimental effect when removed (Metonymic F1 dropped by 0.063), followed by the Constitutive role (−0.053). Removing POS tags and lemmas produced only modest drops (−0.018 and −0.019, respectively), suggesting that surface morphological features play a secondary supporting role relative to the Qualia-based semantic features. The most notable finding is the behaviour of the Named Entity feature: its removal increased Metonymic F1 by 0.054. This counter-intuitive result suggests that NE labels may introduce a classification bias, causing the model to rely on entity type rather than contextual semantic cues. Metonymy in Marathi appears to be more reliably signalled by the surrounding morpho-syntactic and semantic features than by entity category alone.

The Telic role condition produced the highest Macro F1 (69.9%) and Metonymic F1 (70.9%) across all ablation conditions, indicating that removing the Telic role improves performance. Telic role encodes the purpose or function of an entity. This role introduced classification noise among the metonymic expressions in the current dataset. 114

Token	NE Tag	POS	Predicted Label	Contextual Trigger
राष्ट्राध्यक्ष (President)	O	NOUN	Literal	Trump (PER)
यांनी (By / Agentive marker)	O	PRON	Literal	Trump (PER)
हे (This / He)	O	PRON	Metonymic	Zohran (PER)
प्रचार (Campaign)	O	NOUN	Metonymic	Republican Party (ORG)
चर्चचा (Church's)	O	NOUN	Metonymic	institutional context

Table 4: Examples of non-NE marked as literal or metonymic.

non-NE tokens were tagged as either metonymic or literal. Table 4 shows five representative examples of non-NE tokens assigned a literal or metonymic label by the model, along with the contextual NE trigger that likely drove the prediction.

In the first two rows, the presence of a known Person entity (*Trump*) in the immediate context leads the model to assign a Literal label to adjacent non-NE tokens. The model correctly recognises that these tokens are functioning in their literal roles (referring to an actual individual), but their labelling as non-NE items increases the false positive count for the Literal class. The last three rows illustrate that metonymic labels are assigned to nouns and pronouns that are contextually adjacent to an Organisation or Person entity whose presence signals an institutional or representative reading. The pronoun हे (this/he) is labelled Metonymic because it is the pronominal trace of *Zohran*, a PER entity used in a representative context. प्रचार (campaign) is marked Metonymic because it co-occurs with Republican Party, an ORG entity that conventionally activates the Telic and Agentive Qualia roles. Finally, चर्चचा (church's) is labelled Metonymic in an institutional context where the genitive marker implies organisational agency. These examples explain the low preci-

sion. The model's contextual window for Qualia feature activation is wide enough to capture genuine metonymic triggers, but it does not apply strict NE-boundary. Every token inside the influence zone of a metonymically active NE becomes a candidate for labelling, regardless of its PoS category. This is a feature when it enables the model to detect metonymic effects that extend beyond the head entity and a liability when it affects the precision.

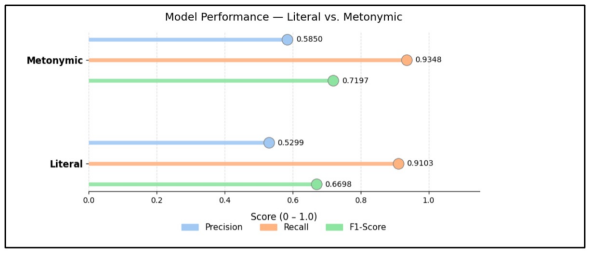


Figure 2: Performance of the model on literal and metonymic categories

5. Limitations

This study has several limitations, which are addressed in ongoing and future research.

- **Corpus size and single annotator:** The corpus size remains restricted at 890 sentences. Additionally, the dataset was annotated solely by a single domain expert; inter-annotator agreement metrics have not yet been established to measure human baseline performance on Marathi metonymy.
- **Noise:** The ablation results indicate that the Telic role introduces noise in the current corpus. Future work will investigate whether a domain-adaptive role assignment, can recover the expected benefit.
- **Integration of case markers and Selectional Preferences:** A particularly promising direction currently under active development by the author but outside the scope of this paper is to formally integrate case markers and Selectional Preference (SP) with Qualia role vectors. Marathi uses several case markers frequently. Such integration would enable the model to jointly reason over morphosyntactic evidence and structured semantic constraints during inference.
- **Named entity label bias:**The finding that removing NE labels improves Metonymic F1 suggests that flat NE type categories introduce classification bias. Future work will explore fine-grained entity tagging and entity-context interaction modelling to address this.

6. Conclusion

This study showed that metonymy detection in Marathi named entities is achievable with small datasets. The hybrid model used deterministic Qualia role vectors derived from Pustejovsky's Generative Lexicon alongside transformer embeddings to detect metonymy in Marathi text. It achieved a high recall for both literal and metonymic cases. Ablation test confirmed that all features are influential, while the NE feature introduced an unexpected classification bias. Although limitations remain, this experiment showed encouraging results. Future work will expand the Marathi metonymy corpus, integrate Selectional Preferences and case markers with Qualia-informed type coercion, and broaden the role assignment schema to improve the performance of the system.

7. Bibliographical References

References

Dan Fass. 1988. An account of coherence, semantic relations, metonymy, and lexical ambiguity resolution. In *Lexical Ambiguity Resolution*, pages 151–177. Elsevier.

Dan Fass, Allan Lesgold, and Vimla Patel. 1997. *Processing metonymy and metaphor*, volume 1. Ablex Publishing Corporation Greenwich, CO.

Saptarshi Ghosh and Tianyu Jiang. 2025. ConMeC: A dataset for metonymy resolution with common nouns. 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 6493–6509, Albuquerque, New Mexico. Association for Computational Linguistics.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources.

H Burcu Kupelioglu, Tankut Acarman, Tassadit Amghar, and Bernard Levrat. 2016. Recognition of metonymy by tagging named entities. *WSEAS Transactions on Computer Research*, 4:81–85.

H Burcu Kupelioglu, Tankut Acarman, Bernard Levrat, and Tassadit Amghar. Helping metonymy recognition and treatment through named entity recognition.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial intelligence*, 135(1-2):145–198.

Katja Markert and Malvina Nissim. 2002a. Metonymy resolution as a classification task. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 204–213.

Katja Markert and Malvina Nissim. 2002b. Towards a corpus annotated for metonymies: the case of location names. In *LREC*.

Katja Markert and Malvina Nissim. 2007. Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41.

Muhammad Elyas Meguellati, Rohana Binti Mahmud, Sameem Binti Abdul Kareem, Assaad Oussama Zeghina, and Younes Saadi. 2022. Feature selection for location metonymy using augmented bag-of-words. *IEEE Access*, 10:81777–81786.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Thierry Poibeau. 2006. Dealing with metonymic readings of named entities. *arXiv preprint cs/0607052*.

James Pustejovsky. 1991. The generative lexicon. *Computational linguistics*, 17(4):409–441.

Contextualising (Im)plausible Events Triggers Figurative Language

Annerose Eichel, Tonmoy Rakshit, Sabine Schulte im Walde
Institute for Natural Language Processing
University of Stuttgart
{annerose.eichel,tonmoy.rakshit,schulte}@ims.uni-stuttgart.de

Abstract

This work explores the connection between (non-)literalness and plausibility at the example of subject-verb-object events in English. We design a systematic setup of plausible and implausible event triples in combination with abstract and concrete constituent categories. Our analysis of human and LLM-generated judgments and example contexts reveals substantial differences between assessments of plausibility. While humans excel at nuanced detection and contextualization of (non-)literal vs. implausible events, LLM results reveal only shallow contextualization patterns with a bias to trade implausibility for non-literal, plausible interpretations.

Keywords: plausibility, (non-)literalness, human vs. LLM generation

1. Introduction and Related Work

How plausible do you judge a situation where the *heat catches a cyclist*? Based on the literal reading of this described situation, one might expect that the majority of human annotators would consider this situation as rather implausible. Yet, previous research has demonstrated that humans are eager to interpret such a seemingly implausible event. In the current study, we investigate to which extent the interpretation of implausible events is driven by the potential to frame them in a non-literal context. For instance, the above situation has been contextualized in figurative example sentences such as *During the intense desert race, the scorching heat caught the cyclist off guard, forcing him to stop for water and shade.* and *The unexpected heat caught the cyclist unaware.*

As the example illustrates, interpreting and distinguishing plausible from implausible events is a crucial and non-trivial building block of natural language. A range of work has explored semantic plausibility using subject-verb-object (svo) events in English leveraging embedding-based neural networks (Wang et al., 2018), transformers (Porada et al., 2019; Emami et al., 2021; Porada et al., 2021), and LLM-based methods (Kauf et al., 2024). However, prior research so far explicitly focused on literal events (Wang et al., 2018) or other aspects of plausibility. Our work addresses this gap and explores the connection between figurative language and plausibility. To do so, we adopt definitions from previous work (Wilks, 1975; Resnik, 1993; Wang et al., 2018) and consider plausibility in a binary setting. *Plausible* events include not only highly typical events but also untypical events (Wilks, 1975), potentially novel events (Wang et al., 2018), and seemingly trivial events such as "a person breathes" that are not necessarily attested in an existing corpus (Gordon and Van Durme, 2013). In comparison, fully implausible events do not allow any semantically valid interpretation; neither a literal nor a figurative reading, for example, through creative metaphors (Griciūtė et al., 2022).

For our study, we rely on a small subset of svo event triples that were previously annotated as (im)plausible (Eichel and Schulte im Walde, 2023). Crucially, the original triples are balanced with regard to the degree of concreteness vs. abstractness of the involved constituent words, because concepts can be described in accordance with the way people perceive them (Barsalou and Wiemer-Hastings, 2005; Brysbaert et al., 2014): Concrete concepts such as *trampoline* can be seen, heard, touched, smelled, or tasted. In contrast, abstract concepts such as *realism* cannot be perceived with the five senses. In between these two extremes on the scale, mid-range concepts such as *punctuality* are situated. The provided abstractness information allows us to connect not only (im)plausibility to (non)literalness but to additionally integrate an interaction with conceptual abstractness, thus implicitly relating to Conceptual Metaphor Theory (Lakoff and Johnson, 1980) as a mapping from abstract to concrete concepts to trigger metaphorical meanings as a special case of non-literal language.

More specifically, we make use of 411 svo triples with plausibility judgments, and ask humans and LLMs to make a binary judgement about their figurative language, plus providing example sentences. Our novel dataset contains a total of 6,497/14,555 judgments and 6,497/3,288 unique sentences generated by humans/LLMs.[1] We use the collected judgements and sentences to compare human and LLM generations. In the context of plausibility, humans have been observed to tend towards sense-making with great nuance and willingness to interpret even whimsical sentences (Griciūtė et al.,

[1] www.github.com/AnneroseEichel/NLE2026

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 93–105
11 May 2026.

2022; Eichel and Schulte im Walde, 2023). Regarding LLMs, while they are equipped for semantic interpretation with (world) knowledge learned through distributional patterns in vast amounts of training data, almost all of their data are plausible. We thus formulate the following research questions:

- RQ1: Does figurative language interact with event plausibility, and how does this interaction relate to the abstractness of the event constituents?
- RQ2: How do human annotations compare to LLM judgments regarding figurative language and event plausibility?
- RQ3: Which qualitative differences can be observed for human vs. model-produced contextualizations of (non-)literal implausible events?

Our contributions are threefold: (i) We present a collection of judgments and example contexts from humans and four LLMs, using a systematic setup of plausible and implausible event triples in combination with abstract and concrete constituent categories. (ii) We provide detailed insights into human vs. LLM judgments for predicting (non-)literalness, and (iii) we conduct a careful analysis of human- and model-generated contexts as well as repair mechanisms for seemingly implausible events.

2. Data

We use the plausibility dataset PAP (Eichel and Schulte im Walde, 2023). PAP encompasses a balanced set of 1,733 subject-verb-object triples in English extracted from Wikipedia (originally plausible) and automatically perturbed triples (originally implausible). All events are labeled by each component's concreteness ranging from abstract (a), over mid-range (m), to concrete (c) (Brysbaert et al., 2014). PAP is balanced across all possible combinations of abstractness such as events consisting of only highly concrete words such as "*person calls town*" (ccc) or fully mixed events such as "*career reestablishes chicken*" (amc). Triples are annotated through crowd-sourcing with subjective assessments of plausibility on a degree scale (1–5) ranging from implausible to plausible. PAP ratings include raw annotations as well as original plausibility labels, and provide clear majority-based ($\geq 70\%$) aggregations. For this study, we use a subset satisfying the following criteria: across all abstractness combinations, we draw a random sample of event triples where (i) original and human-annotated majority label correspond to each other such as "*album breaks genre*" (orig.: *plausible*; PAP maj.: *plausible*), and (ii) original and human-annotated majority label differ such as "*collection needs autonomy*" (orig.: *implausible*; PAP maj.: *plausible*). An overview is shown in App. A.1, Table 4.

3. Methods

For each svo event in our dataset sample, we collect judgments and example sentences from both humans and LLMs. We ask humans to (1) select a label for whether an event is figurative, literal, or neither, based on the combination of component meanings, and (2) produce an example contextualizing the event (only if figurative or literal), or a sentence contextualizing an altered event if neither (cf. App. A.2, Figure 3 for annotation instructions). Then, we prompt LLMs to complete the same tasks.

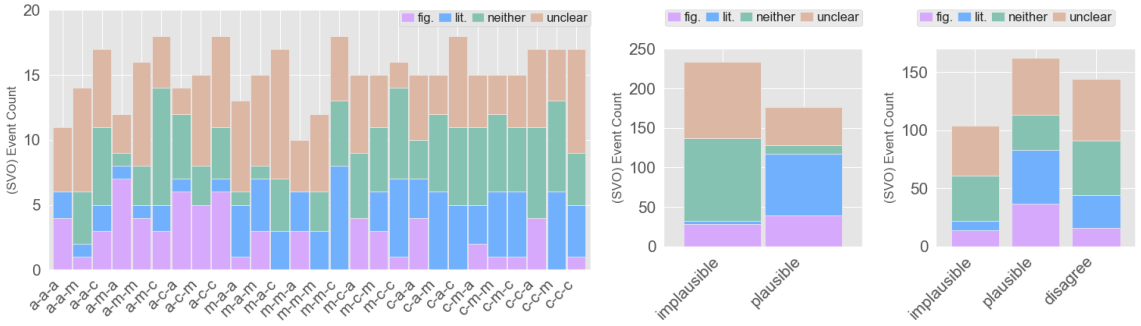
Human Annotation We use Prolific and Google Forms as study tools. Participants are required to reside in the UK, US, Ireland, or Australia and hold corresponding citizenship, speak English as their primary language, and have a Prolific approval rate of $\geq 98\%$. Items are shown in batches of $\approx$25 items with one target shown per page. For each item, we collect $16^{\pm 1.45}$ responses from a total of 240 annotators. We make sure that each annotator contributes $< 1\%$. to the collection. Our annotator sample has a median age of 38 years, is slightly skewed towards female (56.7%) over male annotators (43.3%), and resides mainly in the US (57%). For full demographic details, cf. App. A.2. Across abstractness combinations, we obtain a total of 6,497 judgments and example sentences.

Modeling For model predictions, we use the Instruct versions of four LLMs. We focus on moderate parameter count and test four multilingual model families: Qwen3-4B (Qwen Team, 2025), Gemma3-4B (Gemma Team, 2025), Mistral-7B (Jiang et al., 2023), and Llama3.1-8B (Grattafiori et al., 2024). The LLM prompts for label and text generation are based on instructions for humans with (few-shot) and without (zero-shot) examples (cf. App. A.3, Figure 6 and Figure 5 for prompts). (1) For label prediction, we aggregate results across five model runs with different random seeds and three prompts with varying phrasing and output formats. (2) For example sentence generation, we replicate human instructions as closely as possible and obtain generations for one seed. All prompting is performed with model default settings and a 64 token limit.

4. Results

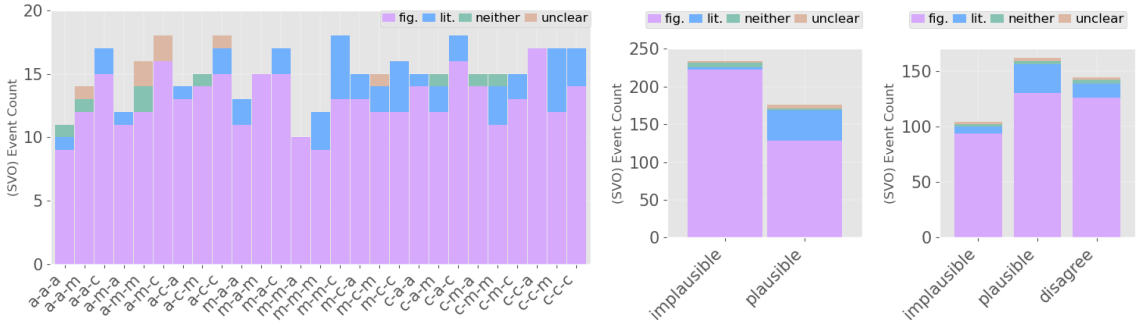
4.1. Figuratively, literally, unclear, or actually not plausible at all?

To explore RQ1, i.e., how figurative language interacts with event plausibility and event constituent abstractness, we first provide a deep-dive into human judgments. To shed light on RQ2, we then assess how human annotations compare to SOTA LLMs' judgments.



(a) **Human** judgements across abstractness comb. (b) orig. label (c) PAP rating

Figure 1: Analysis of **human** figurative majority-assigned labels assigned across (a) 27 **abstractness combinations** ranging from most abstract on the left to most concrete one the right, (b) in comparison with **original labels** used to create PAP, and (c) **PAP majority ratings**.



(a) `Qwen3-4B` judgements across abstractness comb. (b) orig. label (c) PAP rating

Figure 2: Analysis of `Qwen3-4B` figurative labels (zero-shot), similarly to Figure 1.

Human Judgments To investigate the relationship between **abstractness** as well as original and PAP **plausibility** ratings and figurative meanings of the targets, we visualize the number of svo events judged by the majority of participants. Here, majority is defined as $\geq 60\%$. Whenever no majority is reached, we assign the label *unclear*. Figure 1 presents three perspectives on the distribution of judgments. Across plots, label categories are: *figurative* (violet), *literal* (blue), *neither* (no contextualization is possible, green), and *unclear* (orange).

We first look at the interplay between the **abstractness of the svo events, and their perception as figurative vs. literal language**. Across the x-axis, we observe a clear trend. More concrete events (e.g., (*ccc*), (*cam*), (*cac*)) are judged more literally, while more abstract events (e.g., (*aaa*), (*ama*), (*mca*)) are judged more figuratively. More specifically, subject and object abstractness exert greater influence on (non-)literalness. In particular, concrete or mid-scale verbs in such events lead to predominantly literal readings of events, confirming prior work (Khaliq et al., 2024; Knupleš et al., 2026). In turn, we observe a limited influence of verb abstractness. Items for which neither a figurative nor a literal reading or a context could be inferred such as "*payload lives blowout*" or "*city folds fruit*" accumulate on the more concrete end of the scale. Here, concrete objects strongly influence events that are otherwise abstract.

Next, we focus on the relationship between **plausibility and figurative language**. We consider both the original labels underlying the PAP dataset, and the majority PAP ratings. Originally plausible items such as "advance guarantees freedom" are judged more figuratively than originally implausible items such as "*copy inflates disbelief*". While the difference is rather small for original labels, a larger disparity is observed for PAP ratings. When inspecting targets judged literally such as "*owner secures trademark*", a virtual clear-cut between plausible and implausible items emerges. While both original and majority PAP plausible items are judged mainly literally, implausible items are virtually never perceived as literal.

Following our qualitative inspection, we further **quantify whether assigned (non-)literal labels are related with event abstractness, original label, or PAP ratings** using χ^2 tests of independence (Pearson, 1991). Strength of association is examined with Cramér's V (Cramér, 1999). An overview of results is presented in Table 1. We find significant associations ($p < .001$) between abstractness combinations and figurative/literal labels, with a mod-

		Human		Qwen3-4B		Llama3.1-8B		Mistral-7B		Gemma3-4B	
Figurative	df	χ^2	V	χ^2	V	χ^2	V	χ^2	V	χ^2	V
ABSTRACTNESS	26	**63.64*** **	0.39	23.10	0.24	48.13**	0.34	41.50*	0.32	42.82*	0.32
ORIG. LABEL	1	6.91**	0.13	37.68	0.30	0.17	0.02	4.64*	0.11	33.10	0.28
PAP RATING	3	13.49**	0.18	6.20	0.12	2.88	0.08	1.59	0.06	7.68	0.14
Literal											
ABSTRACTNESS	26	43.93**	0.33	34.81	0.29	50.88**	0.35	42.67*	0.32	43.75*	0.33
ORIG. LABEL	1	**111.33*** **	0.52	45.72	0.33	0.46	0.03	4.66*	0.11	32.23	0.28
PAP RATING	3	**17.29*** **	0.21	7.85*	0.14	2.25	0.07	5.42	0.11	8.09*	0.14
Neither											
ABSTRACTNESS	26	32.80	0.28	27.33	0.26	23.33	0.24	-	-	32.86	0.28
ORIG. LABEL	1	71.96	0.42	0.45	0.03	1.53	0.06	-	-	1.19	0.05
PAP RATING	3	13.73**	0.18	0.04	0.01	0.48	0.03	-	-	2.16	0.07
Unclear											
ABSTRACTNESS	26	29.59	0.27	32.56	0.28	28.31	0.26	39.61*	0.31	40.18*	0.31
ORIG. LABEL	1	8.23**	0.14	1.33	0.06	0.80	0.04	0.32	0.03	1.50	0.06
PAP RATING	3	4.17	0.10	0.15	0.02	2.10	0.07	0.59	0.04	5.42	0.11

Table 1: Associations between figurative language and abstractness, original label, or PAP ratings. χ^2 indicates *significance* ($p < .05$:**, $p < .01$:**, $p < .001$: ***) and Cramér's V measures *strength* of association. Model results are based on zero-shot prompts.

erate effect size. This finding further underlines previous work (Khaliq et al., 2024; Knupleš et al., 2026) focusing on verb-object (v,o) pairs where they find an increase in figurative majority-based judgments predominantly influenced by object abstractness. We further find a significant association ($p < .001$) between original and literal labels, with a strong effect size. For literal labels and majority PAP, we also observe a statistically reliable association ($p < .001$) albeit with a weaker effect size.

Human vs. LLM Judgments We compare majority-assigned label distributions obtained by humans vs. four SOTA LLMs where majority is defined as $\geq 60\%$. Results across models and prompt types are shown in Table 2 with performance equally low for all four models, i.e., they mostly disagree with human judgments. When visualizing majority-assigned labels, we observe clear differences across models and prompt types. For reasons of space, we illustrate model results at the example of Qwen3-4B results in Figure 2 and provide a full overview in App. B, Figure 7.

Zero-shot prompts trigger Gemma, Qwen, and Mistral to assign overwhelming majorities of plausible, and specifically figurative readings. This holds for both originally plausible and implausible events. A notable exception is Llama which assigns significantly more literal interpretations across original labels and abstractness combinations. This trend is only partially observable for the other three models which assign literal readings for more concrete events. In comparison to the label distribution based on human annotations, there is a notable absence of implausible instances across all models. In particular, Mistral does not produce a single majority assignment for *implausible*. Similarly to the analysis of human judgements, we conduct a quantitative analysis to explore the relationship between figurative language and abstractness, original label, and PAP rating. Results are shown in Table 1, indicating associations ($p < [0.01, 0.05]$) with moderate effect size between *figurative* and *literal* labels and abstractness for all models except Qwen.

Few-shot prompts (cf. App. B, Figure 8) change Gemma results with a significant increase in implausible and unclear events. Interestingly, only marginally different results are observable for both Qwen and Mistral, which could either point to strong prediction stability across prompts or disregard of contextual information in the middle of a prompt. Lastly, Llama results change to overall more figurative events assigned. Additional quantitative inspections (cf. App. B, Table 5) underline the relation between plausible labels and abstractness with stronger associations than for zero-shot prompting.

We further explore for both zero- and few-shot settings **which prompt template leads to the strongest bias towards figurative interpretation** of the examined events. Results are shown in App. B, Table 6, indicating that prompt templates based on human instructions introduce the least bias across models.

In summary, our hypotheses are confirmed for **human-annotated** events: (i) The more concrete svo event constituents are, the more likely contextualization fails, i. e., events being judged as neither figuratively nor literally meaningful, but implausible (nonsensical). This finding underlines previous work on the influence of event abstractness

MODEL	ZERO-SHOT		FEW-SHOT	
	ACC.	τ	ACC.	τ
Gemma3-4B	0.27	0.09	0.30	0.01
Qwen3-4B	0.27	**-0.15**	0.25	**-0.16**
Mistral-7B	0.20	-0.10	0.21	-0.05
Llama3.1-8B	0.28	0.04	0.32	0.05

Table 2: Model performance for label prediction across prompt types. Predictions are aggregated across prompt templates and five model runs. We report *accuracy* (acc.) and *Kendall's* τ using human majority-assigned decisions as reference value. Bold: $p < 0.001$

on plausibility (Eichel and Schulte im Walde, 2023) and complements research on semantically anomalous vs. truly nonsensical expressions (Olsen and Padó, 2026). (ii) Confirming our hypothesis, the more abstract svo event constituents are, the more frequently plausibility is perceived, and the more probable is a figurative reading. In contrast, **LLM-predicted** results deviate from our hypothesis as we find (i+ii) a strong bias for plausibility, and specifically figurative language across categories for Qwen, Mistral and Gemma, while overall, Llama results are closer to human judgments. However, human-annotated results are only weakly mirrored with models trading implausibility for plausibility.

4.2. Qualitative Characteristics of Human- vs. LLM-Produced Contexts

We qualitatively evaluate generated contexts to assess how humans vs. models contextualize (im)plausible svo events. Across our 27 abstractness combinations, we sample up to four examples (one per label) produced by humans or models (zero-shot), yielding 97 contexts. We sample 3 example sentences per investigated event from human-generated contexts. We label one model generation per event. Events incorrectly (not) containing the original svo event are labeled *none*. We also assign *none* in case of more than one changed constituent or in case of constituent changes despite a plausible judgement. Whenever events correctly contain the original event but are semantically invalid, we assign the label *anomalous*. We follow Olsen and Padó (2026)'s labeling scheme and annotate contexts as *specific* if no self-reported indication of generic settings such as *fantasy* story is present in produced example contexts. In case of changes, we track altered event constituents.

Results for human-produced and model-generated contexts are reported in Table 3, highlighting a substantial number of specific contexts by both humans and LLMs. In comparison to humans, LLMs rarely predicted the label *neither* in which case an event constituent should be altered to enable contextualization. Nevertheless,

	H1	H2	H3	LL	QW	MI.	GE
Specific	94	92	93	51	52	27	17
Altered (s)	10	8	8	-	-	-	2
Altered (v)	6	13	13	-	-	-	-
Altered (o)	15	8	8	-	-	-	-
Anomalous	-	1	2	6	11	1	22
None	3	4	2	40	34	69	58

Table 3: Human (H) vs. LLM (LL: Llama, QW: Qwen, MI: Mistral, GE: Gemma) context patterns.

especially Mistral and Gemma frequently alter events despite a predicted figurative or literal label. Moreover, in the few cases were *neither* was assigned, models mostly fail to correctly change only one constituent *and* produce a valid sentence. Further, LLM contextualization strongly adheres to original event syntax, as highlighted by contexts to the originally and majority PAP implausible event *"license hinders ice"*. Qwen repeats the event (*"The event license hinders ice."*). Both Llama and Gemma add a single object (*"The event license hinders ice skating."*). Mistral's generation illustrates a common failure across all models: incorrect substitution despite a plausible judgement (*"The ice sculpture exhibition was hindered by the event license restrictions.")*

In comparison, human-produced contexts are based on a majority-assigned *neither* label with examples altering the subject (*"The recent sunny weather hinders ice for hockey player."*) or the object (*"The strict license hindered access to the restricted research facility."*) or providing an actually supporting, non-anomalous context (*"A license doesn't have the capability to hinder ice from forming."*) While humans use a wide vocabulary range and vary syntax to generate meaningful contexts, all models over-use the term *event*, and adhere to mostly nouns, verbs, and simple syntactic structures. In conclusion, LLM results reveal shallow patterns when compared with human-generated contexts exhibiting great nuance at assessing plausibility and contextualizing and 'repairing' events.

5. Conclusion

This work explored the connection between plausibility and (non-)literalness at the example of svo events in English. Using a carefully selected section of the PAP dataset (Eichel and Schulte im Walde, 2023), we collected and analyzed human- vs. LLM-generated judgments and examples. Our analysis reveals substantial differences between human and LLM assessments for the examined events. While humans excel at nuanced detection and contextualization of (non-)literal vs. implausible events, LLM results reveal shallow context patterns and a strong bias towards plausibility.

6. Acknowledgements

This research was supported by the Hanns Seidel Foundation's Talent Program (first author) and the DFG Research Grant SCHU 2580/4 *MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*. We also thank the reviewers for their helpful comments and nuanced feedback.

7. Limitations and Ethical Considerations

A first obvious limitation of our work is the sole focus on the English language. We expect results to differ for other languages and encourage work on plausibility, (non-)literalness, and abstractness. In this paper, we present a collection of (non-)literalness judgments and examples sentences collected via crowd-sourcing. We employ control items as well as post-processing to minimize the impact of unreliable annotations on our analyses. Approaches of mitigation could be concentrating on labels with high majorities of one label assigned or use e. g., probabilistic approaches to aggregate labels. We pay participants fairly and seek transparent communication of decisions whenever necessary during the annotation approval process. Furthermore, in our work we use a reasonable set of heuristics to parse LLM-generations. It is possible that more complex approaches might lead to different results based on the parsing process. Finally, we use LLMs that are known to exhibit bias which might be reflected in the way events are judged as well as in the style and content of generated contexts.

8. Bibliographical References

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating Abstract Concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Harald Cramér. 1999. *Mathematical methods of statistics*, volume 9. Princeton university press.

Annerose Eichel and Sabine Schulte im Walde. 2023. A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 31–45, Toronto, Canada. Association for Computational Linguistics.

Ali Emami, Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. ADEPT: An adjective-dependent plausibility task. 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)*, pages 7117–7128, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, page 25–30, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.

Bernadeta Griciūtė, Marc Tanti, and Lucia Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 173–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Mohammed Khaliq, Diego Frassinelli, and Sabine Schulte Im Walde. 2024. Comparison of Image Generation Models for Abstract and Concrete Event Descriptions. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 15–21, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2026. Literally Concrete or Figuratively Abstract? Multilingual Concreteness Norms for Verb-Object Expressions. *Transactions of the Association for Computational Linguistics (TACL)*. In press.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Katrin Olsen and Sebastian Padó. 2026. Finding Sense in Nonsense with Generated Contexts: Perspectives from Humans and Language Models. Manuscript.

Karl Pearson. 1991. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Breakthroughs in Statistics: Methodology and Distribution*, page 11.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 123–129, Hong Kong, China. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1732–1743, Online. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 Technical Report.

Philip Stuart Resnik. 1993. *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

A. Appendix

A.1. Data

Table 4 presents an overview of the target svo events sampled from the PAP dataset.

orig. labels	**PAP ratings**		
	plausible	*disagree*	*implausible*
plausible	81	64	31
implausible	81	80	77

Table 4: Overview of number of target event triples sampled from PAP.

A.2. Human Annotation

Human Annotator Demographics Figure 4 shows an overview of annotator demographics. Subplot (a) visualizes the age distribution of involved annotators. Mean age is 39.5 years and median age is 38 years. Subplot (b) presents the distribution between female and male participants. Please note that this is based on self-reported biological sex of participants. We do not collect information on gender identity. We report annotators' employment status in subplot (c) with the majority of participants either working full-time or part-time. "No paid work" refers to individuals focusing on care work as well as retired or disabled individuals. "Soon new job" denotes participants who start a new job in the next month (which does not mean that they are not employed at the moment they took part in the study). "Unemployed" implies that someone is unemployed *and* job-seeking. "DATA_EXPIRED" refers to long-time participants on Prolific who have not updated their Prolific profile for a longer period of time. Some information such as employment or student status hence might get marked as expired. As visualized in subplot (d), self-reported simplified ethnicity groups are mainly White (67%) and Black (25%). While nationality as shown in subplot (e) needs to include UK, Ireland, U.S., or Australia, participants might have dual citizenship (e.g., the UK allows for that). Subplot (f) lists countries where participants reside.

A.3. Modeling

We present prompt templates for zero- and few-shot prompting in Figure 6 and Figure 6.

B. Results

Human vs. Model Label Predictions We present **zero-shot** model results for predicting figurative labels for all models in Figure 7. **Few-shot**

Guidelines

You will be given 30 three-word events such as "cat eat sardine" or "friend grasp meaning".

Tasks:

1. Decide whether the event description is clearly based on the meanings of the three words.
For example**, the event "cat eat sardine" is literally describing the event of a cat eating a sardine**. In contrast, **the event "friend grasp meaning" does not describe a friend literally grasping a meaning: the event meaning is figurative,** rather than literal.

2. Together with your decision regarding whether an event is literal or figurative, we also ask you to **provide an example sentence** including the three-word event. For example:

- Literal usage: *I saw a cat eating a sardine near the lake today.*
- Figurative usage: *My friend quickly grasped the meaning of the mathematical problem.*

3. In some cases you will neither be able to identify a literal nor a figurative meaning, because the event is absolutely implausible, and it is not possible to come up with any interpretation at all. An example of such an **absolutely implausible event is "plea overrun rain"**. In this case, we ask you to alter one of the three words and then **provide an example sentence**. For instance, changing "*overrun*" to "*summons*" forms the phrase "**plea summons rain**".

- Adjusted event: *The farmer's plea summons the rain to save the crops.*

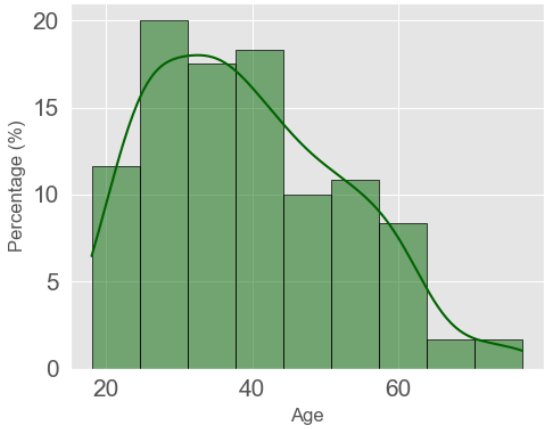
Decision for "reign spreads power" *

- ○ literal
- ○ figurative
- ○ neither: the event is implausible

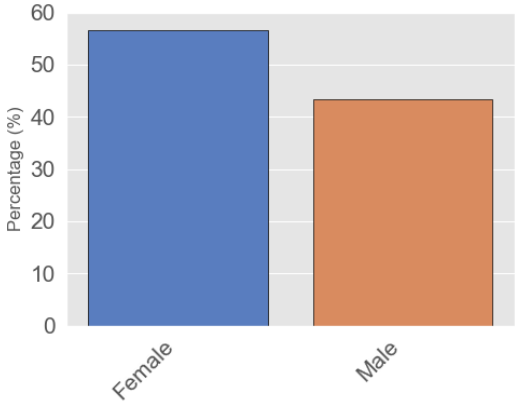
Example sentence for "reign spreads power" (or an event variation, if this event is absolutely implausible)

Your answer

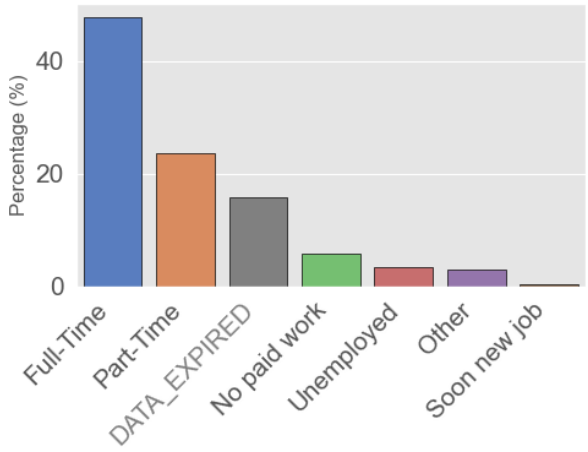
Figure 3: Annotation interface



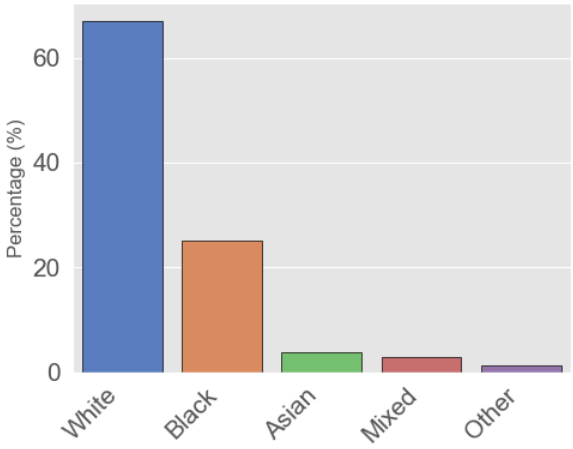
(a) Age.



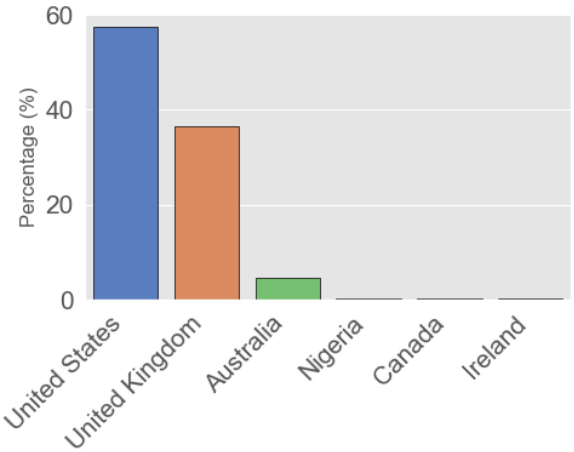
(b) Female and male (sex; gender identity information not collected).



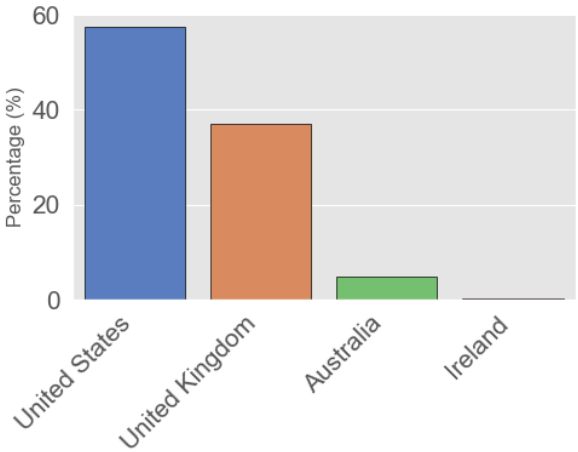
(c) Employment status.



(d) Self-reported simplified ethnicity groups.



(e) Nationalities (dual citizenship possible).



(f) Country of residence.

Figure 4: Distributions of annotator demographic features.

model results for predicting figurative labels are presented for all models in Figure 8.

Further, quantitative analysis results for few-shot modeling results are presented in Table 5. We include values from human analysis for reference.

We analyze which prompt template leads to strongest bias towards figurative interpretation. We consider both zero- and few-shot prompt templates and show results in Table 6.

```
Is the following event figurative or literal or neither?
Event: {event}
Answer with only one label: figurative or literal or neither.
Respond in the following format:
Label: <figurative|literal|neither>
```

```
Determine whether the event below is figurative or literal or neither.
Event: {event}
Respond in the following format:
Label: <figurative|literal|neither>
```

```
Decide whether the event description is clearly based on the meanings of the three
↪  words.
Event: {event}
Answer with only one label: <figurative|literal|neither>.
Decision: <label>
```

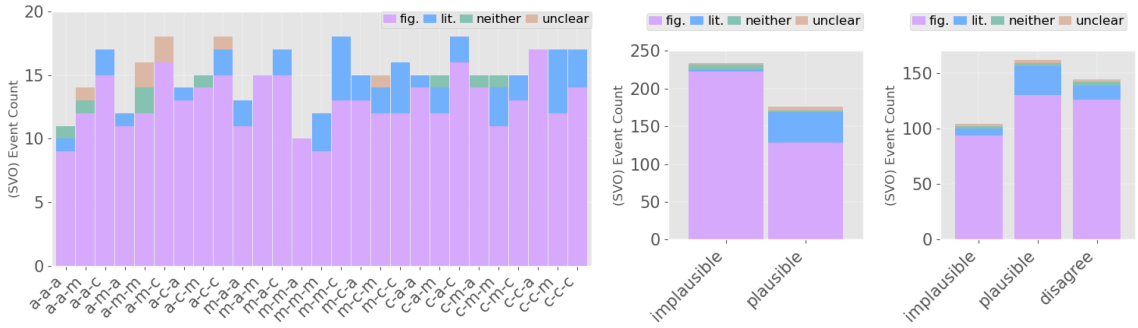
Figure 5: **Zero shot** prompt templates, from top to bottom including the task formulation as a *question* (top), a statement (middle), and as closely as possible based on the instruction for human annotation but condensed in one sentence (bottom).

```
Decide whether the event description is clearly based on the meanings of the three
↪  words.
For example, the event "cat eat sardine" is literally describing the event of a cat
↪  eating a sardine.
In contrast, the event "friend grasp meaning" does not describe a friend literally
↪  grasping a meaning:
the event meaning is figurative, rather than literal.
Event: {event}
Answer with only one label: <figurative|literal|neither>
Decision: <label>
```

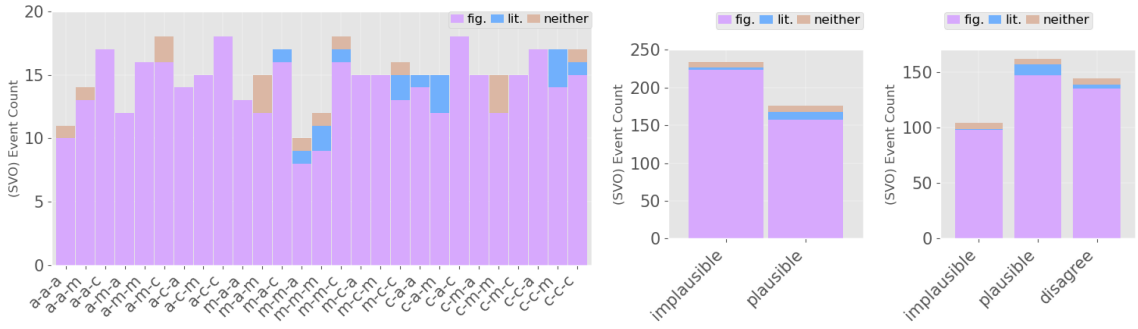
Figure 6: **Few shot** prompt template. The prompt is based as closely as possible on the instruction for human annotation while optimizing for the shortest length.

		Human		**Qwen3-4B**		**Llama3.1-8B**		**Mistral-7B**		**Gemma3-4B**	
Figurative	df	χ^2	V	χ^2	V	χ^2	V	χ^2	V	χ^2	V
ABSTRACTNESS	26	**63.64*** **	0.39	43.66*	0.33	46.71**	0.34	51.37**	0.35	53.59**	0.36
ORIG. LABEL	1	6.91**	0.13	41.05	0.32	7.74**	0.14	4.64*	0.11	**13.56*** **	0.18
PAP_RATING	3	13.49**	0.18	15.48**	0.19	4.04	0.10	3.37	0.09	1.23	0.05
Literal											
ABSTRACTNESS	26	43.93**	0.33	50.84**	0.35	**61.28*** **	0.39	34.08	0.29	**57.32*** **	0.37
ORIG. LABEL	1	**111.33*** **	0.52	40.70	0.32	**25.35*** **	0.25	10.02**	0.16	**19.26*** **	0.22
PAP_RATING	3	**17.29*** **	0.21	15.73**	0.2	7.37	0.13	3.91	0.10	1.97	0.07
Neither											
ABSTRACTNESS	26	32.80	0.28	33.25	0.28	23.17	0.24	33.14	0.28	29.15	0.27
ORIG. LABEL	1	71.96	0.42	0.02	0.01	0	0	-	-	6.45*	0.13
PAP_RATING	3	13.73**	0.18	1.54	0.06	1.86	0.07	6.54	0.13	8.39*	0.14
Unclear											
ABSTRACTNESS	26	29.59	0.27	28.55	0.26	20.12	0.22	51.42**	0.35	39.34*	0.31
ORIG. LABEL	1	8.23**	0.14	0	0	1.02	0.05	0	0	0.87	0.05
PAP_RATING	3	4.17	0.10	3.73	0.10	10.21*	0.16	1.89	0.07	9.44**	0.15

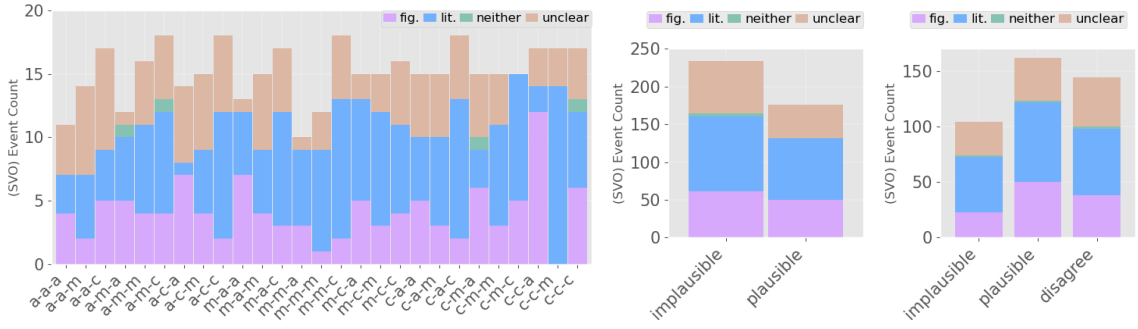
Table 5: Associations between figurative language and abstractness, original label, or PAP ratings. χ^2 indicates *significance* ($p < 0.05$:**, $p < 0.01$:**, $p < 0.001$: ***) and Cramér's V measures *strength* of association. Model results are based on **few-shot** prompts.



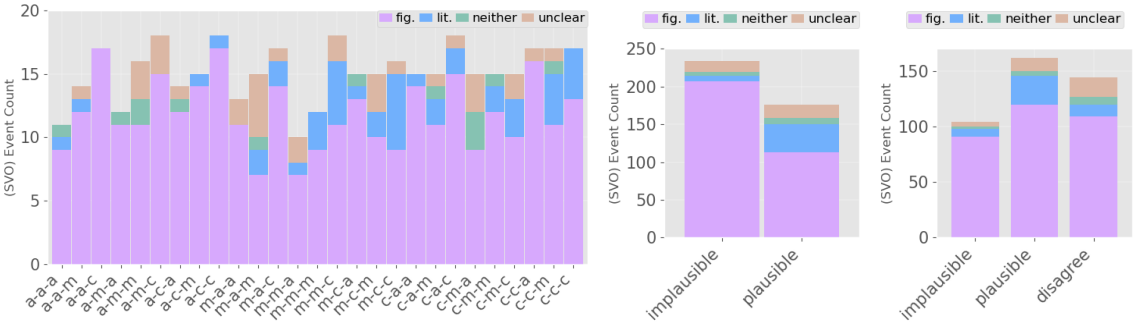
(1a) **`Qwen3-4B`** judgements across abstractness comb. (1b) orig. label (1c) PAP rating



(2a) **`Mistral-7B`** judgements across abstractness comb. (2b) orig. label (2c) PAP rating

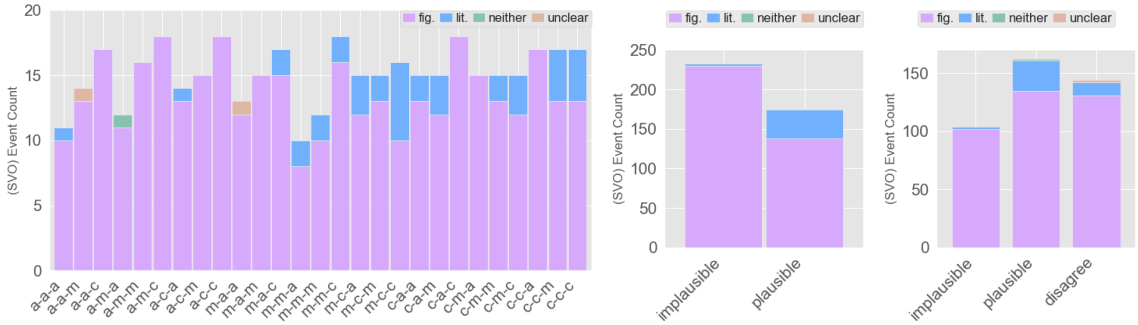


(3a) **`Llama3.1-8B`** judgements across abstract. comb. (3b) orig. label (3c) PAP rating

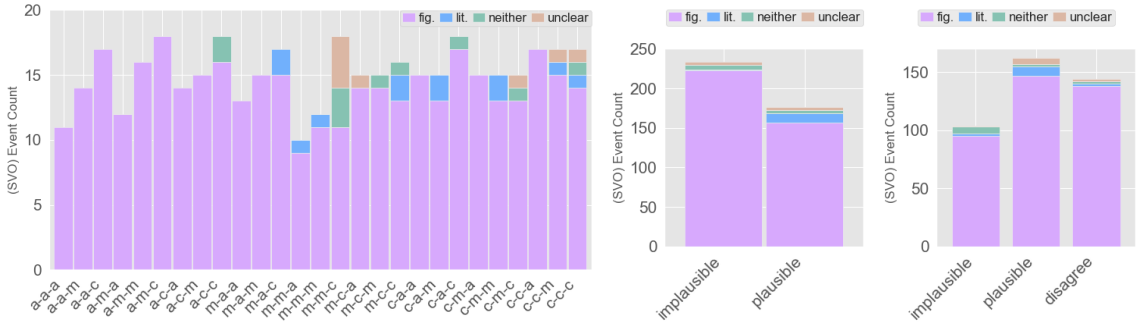


(4a) **`Gemma3-4B`** judgements across abstract. comb. (4b) orig. label (4c) PAP rating

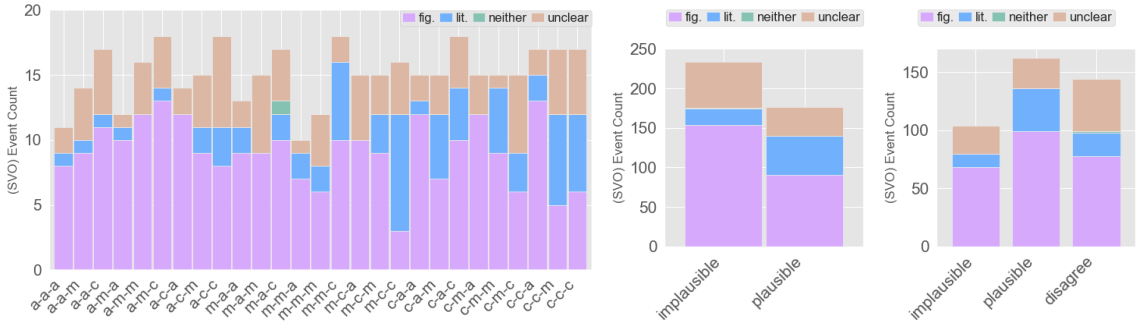
Figure 7: Overview of analysis of **`Qwen3-4B`**, **`Mistral-7B`**, **`Llama3.1-8B`**, and **`Gemma3-4B`** figurative labels (**zero-shot**), similarly to human analysis in Figure 1.



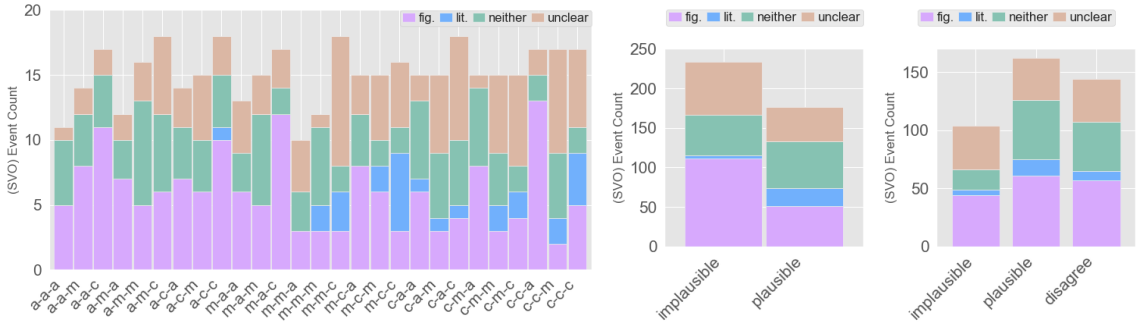
(1a) **`Qwen3-4B`** judgements across abstractness comb. (1b) orig. label (1c) PAP rating



(2a) **`Mistral-7B`** judgements across abstractness comb. (2b) orig. label (2c) PAP rating



(3a) **`Llama3.1-8B`** judgements across abstract. comb. (3b) orig. label (3c) PAP rating



(4a) **`Gemma3-4B`** judgements across abstract. comb. (4b) orig. label (4c) PAP rating

Figure 8: Overview of analysis of **`Qwen3-4B`**, **`Mistral-7B`**, **`Llama3.1-8B`**, and **`Gemma3-4B`** figurative labels (**few-shot**), similarly to human analysis in Figure 1.

	Zero-shot			Few-shot
	question	statement	human instr.	human instr.
Gemma3-4B	67.88	89.29	61.05	22.83
Qwen3-4B	91.00	83.94	63.50	73.90
Mistral-7B	84.91	95.83	78.47	94.51
Llama3.1-8B	23.88	31.39	31.03	61.37

Table 6: Overview of share of only figurative labels per prompt, in percent. Note that percentages do not add up to 100% but each number is a share of 100% labels distributed among *figurative* (shown here), *literal*, *neither*, and *unclear* (not shown). For reference, prompt templates are listed in Figure 6 and Figure 5.

A Novel Dataset and Three Ways to Approach Automatic Metaphor Detection in German Religious Online Forums

Sebastian Reimann, Tatjana Scheffler
Ruhr University Bochum
Department for German Language and Literature
{firstname.lastname}@rub.de

Abstract

In recent years, automatic metaphor detection has received considerable attention within NLP. However, the largest share of research, including most datasets annotated for metaphor, has concentrated on English and a limited set of genres. Automatic metaphor detection for a genre like religious online communication, which is particularly rich in metaphor, remains understudied, in particular since annotated data for this genre is lacking in the first place. This paper aims to close these gaps by offering a novel dataset of posts from German online forums annotated for metaphor, which opens up new research opportunities for automatic metaphor detection for German. Moreover, we present an in-depth exploration in which we evaluate the suitability of different strategies to overcome the relative lack of training data for this task by comparing cross-lingual and cross-genre transfer strategies with the use of LLM prompting. We find that fine-tuning encoder-only language models outperforms the prompting-based approach, that different architectures based on contextual embeddings indeed exhibit considerable differences in their behavior and that smaller in-genre data may be preferable for certain use cases over fine-tuning on larger datasets from different genres.

Keywords: metaphor, figurative language, cross-lingual transfer, online forums, social media

1. Introduction

Metaphoric language is more than just a stylistic tool. It rather points to underlying mappings between semantic domains that structure the way humans perceive the world (Lakoff and Johnson, 1980). For some genres, metaphor is of particular importance, since it serves to describe what otherwise cannot be put into words. For example, in religious communication, transcendent and supernatural entities like God must be described using metaphors from the physical world, such as through describing God as a *father* and believers as *children* (Krech et al., 2023). Religious communication (together with other specialized discourse such as science communication) consequently represents a particularly fruitful use case for the application of automatic metaphor detection methods. Such solutions would also enable potentially new interdisciplinary and quantitative perspectives for the study of religion. However, while automatic metaphor detection has indeed received more and more attention in recent years, the best performing approaches (Choi et al., 2021; Babieno et al., 2022; Li et al., 2023) were almost exclusively applied in scenarios where large amounts of training data that matched the test data in terms of both language and genre were available; particularly focusing on the VU Amsterdam Metaphor Corpus (Steen et al., 2010) (VUAMC), which consists of texts from news, everyday conversation, light fiction and academic discourse. The application to datasets from other genres, as well as cross-lingual applications that involving testing on data in other languages than English, remains scarce. Particularly for German, there is a lack of resources. Existing metaphor datasets for German are either very small and machine-translated versions of English datasets (Berger et al., 2024), not yet publicly available (Egg and Kordoni, 2023) or focus on one specific word class such as particle verbs only (Köper and Schulte im Walde, 2016). Additionally, the study of Reimann and Scheffler (2024) is the only work applying metaphor detection to religious texts.

More recently, there has been a shift from supervised finetuning to zero- and few-shot prompting of generative Large Language Models (LLMs) for many NLP tasks. In theory, this presents new opportunities in cases where data is scarce. However, metaphor detection appears to be a task that still poses problems for generative LLMs (Chen et al., 2024; Liang et al., 2025; Reimann and Scheffler, 2025a,b), which raises questions on how to deal with the data shortage for our specific use case of metaphor detection in Christian online forums. Can LLM prompting be a viable option for scenarios where training data does not match the target test data in language and/or genre or are supervised transfer learning scenarios the better option?

More specifically, to answer the aforementioned issue, we make the following contributions in this paper:

1. We make a novel dataset obtained from German Christian online forums publicly available[1].

[1] https://github.com/SFB-1475/

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 106–118
11 May 2026.

This dataset is, to the best of our knowledge, the largest publicly available German dataset annotated for metaphor and will thus counteract the lack of high-quality datasets for this particular language and genre combination.

2. We explore cross-lingual transfer potentials for German metaphor detection with multilingual, encoder-only language models and English as a source language. We find that such transfer indeed works if the underlying cross-domain mapping is common in both languages.

3. We share the first study that explicitly contrasts cross-lingual, cross-genre transfer and prompting based approaches for automatic metaphor detection. In a supervised setting, we find that using data from similar genres but different languages may indeed be preferred, even if the datasets are smaller and cross-lingual transfer is necessary.

2. Previous Work

In recent years, the most successful body of research on automatic metaphor detection (Mao et al., 2019; Choi et al., 2021; Babieno et al., 2022; Zhang and Liu, 2023) used contextual word embeddings to model the procedures for manual metaphor identification, in particular the Metaphor Identification Procedure (Pragglejaz Group, 2007) (MIP) concerned with the contrast between the meaning of a word in its context and its more concrete basic meaning, and the Selectional Preference Violation (SPV) approach (Wilks, 1975) concerned with the semantic difference between a word and its context. However, these methods for automatic metaphor detection focused almost exclusively on English.

The earliest cross-lingual approach was put forward by Tsvetkov et al. (2014) who use vectors that incorporate features for abstractness, imageability and supersenses in order to detect metaphorical subject-verb-object and adjective-noun constructions with a random forest classifier. Non-English examples are translated before the are vectorized. The F1-scores achieved by their approach remain relatively constant across all four languages (English, Spanish, Russian and Farsi), demonstrating potential for cross-lingual transfer.

Despite the promising early results of Tsvetkov et al. (2014), cross-lingual metaphor detection has been relatively underexplored until recently. Besides several other probing experiments, Aghazadeh et al. (2022) tested the cross-lingual transfer capabilities of pretrained language models for metaphor detection. For these cross-lingual experiments, they used data from the multilingual LCC Corpus (Mohler et al., 2016) containing sentences in English, Spanish, Russian and Farsi. For each language, they finetuned XLM-RoBERTa (XLM-R) on equal portions of the corpus and tested them on data from all languages in the corpus. Although training and testing on the same language always yielded the highest accuracy, for all cross-lingual transfer settings, the models outperformed a random baseline, suggesting that some transfer occurred successfully.

Sanchez-Bayona and Agerri (2022) present CoMeta, a large Spanish corpus annotated via MIPVU, the same guidelines used for the annotation of the VUAMC. They carry out cross-lingual metaphor detection experiments, where they finetune the most successful models from monolingual experiments (XLM-R for the evaluation on Spanish and DeBERTa for the evaluation on English) in a cross-lingual transfer setting, from Spanish to English and vice versa. Training on Spanish and evaluating on English expectedly underperformed the monolingual results. However, finetuning on English and evaluating on Spanish even outperformed the monolingual Spanish setting, which they attribute to the larger size of the English training set compared to its Spanish counterpart.

For metaphor detection for Slovene, Klemen and Robnik-Šikonja (2023) also included a series of cross-lingual experiments, besides monolingual experiments, where they train on either English data from the VUAMC or a combination of English and Slovene data and test on Slovene, using XLM-R and a trilingual model pretrained on English, Slovene and Croatian data (Ulčar and Robnik-Šikonja, 2020). However, they find that English data only had a minor effect on performance.

Berger et al. (2024) used a version of the English metaphor corpus of Gordon et al. (2015) that was translated into German and reannotated via MIPVU to explore the cross-lingual transfer for metaphor detection. They approached the task in two ways: as a sequence classification task providing one label for each word in a sentence and a classification task providing a label for the entire sentence on whether it contains a metaphoric word. In all experiments, they used VUA for training and evaluated both the original English data from Gordon et al. (2015) and their German translations. For the sequence classification task the performance was low overall. For sentence classification, they report more positive results, with SBERT only underperforming a monolingual evaluation by 5 points in F1 (66% vs 61&).

Hülsing and Schulte Im Walde (2024) apply cross-lingual transfer to verb metaphor detection in low-resource scenarios with English as a source language and German (using the particle verb data of Köper and Schulte im Walde (2016)), Latin and

C04-Cross-Lingual-MD

Russian as target languages. They train multilingual BERT in both a zero-shot setting with only English data involved in training as well as in a few-shot setting with 20 target language examples and with the multilingual adapter MAD-X on English and evaluate its performance on German, Latin and Russian. They report mostly satisfactory results, with F1-scores ranging from 60% to 87%, with the best performances on Russian with the MAD-X adapter. Introducing the target language examples, surprisingly, did not bring any improvement.

3. Data

3.1. English

For our English data, we rely on preexisting datasets annotated for metaphor via MIPVU. In order to have a dataset that corresponds in terms of genre with our German data, we use the data of Reimann and Scheffler (2024), consisting of posts from two Christian subreddits.

Additionally, for the comparison of cross-lingual and cross-genre transfer, we include a large dataset from a different genre. Here, the logical choice is the version of the VUA corpus (Steen et al., 2010) used in the 2020 Metaphor Detection Shared Task (Leong et al., 2020). Both the Reddit dataset of Reimann and Scheffler (2024) and the version of VUA used in Leong et al. (2020) only included content words (nouns, verbs, adjectives and adverbs) in their metaphor annotations. The VUA corpus is available under a CC BY-SA 3.0 license and the Reddit data is available under the CC-BY-4.0 license.

3.2. German

For our German dataset, we searched the religious German online forum *jesus.de* for threads where the word *Metapher* ("metaphor") is mentioned, in order to capture as many examples as possible that contain metaphors that were deliberately used as such.

We annotated our data following the MIPVU protocol (Steen et al., 2010), in particular its variant for German (Herrmann et al., 2019), which makes additional recommendations on how to deal with German morphology and the most suitable German dictionaries (Duden and DWDS). We only annotate nouns, verbs, adjectives and adverbs. For its basic procedure, MIPVU asks the annotator to perform the following steps:

1. Read and understand the text
2. Divide the text into lexical units
3. On a word by word basis, define the contextual meaning (3a), decide if a more basic (i.e., more concrete, human-oriented) meaning exists in a corpus-based dictionary (3b), and decide if the two meanings are sufficiently distinct but still related by some sort of similarity (3c)

If these conditions are met, the word is considered to be a *indirect Metaphor-Related Word (MRW)*. Additionally, MIPVU covers so-called *direct* MRWs, which are metaphoric words that are part of a metaphoric comparison (e.g., "strong like a lion"). Steen et al. (2010) argue that they are used in their literal meaning and thus not covered by the standard procedure but they still express a mapping between two domains.

Four annotators were involved in the annotation process, a PhD student of computational linguistics and a PhD student of religious sciences, as well as two student assistants, one with a background in philosophy and English linguistics and one with a background in religious sciences. All annotators are native speakers of German. Initially, we carried out an annotation round to familiarize ourselves with the data as well as the annotation guidelines with all annotators and extensively discussed all cases of disagreement in the curation process.

One aspect that caused initial confusion among annotators were frequently occurring lexicalized expressions in German such as *es gibt* ("there is", lit. "it gives") and *es geht mir gut* ("I'm doing fine", lit. "it goes me fine"), which convey some sense of figurativeness and may possibly be related by metaphor when looking at them through the lens of historical linguistics. However, MIPVU emphasizes to assume a synchronic, modern-day perspective. From this, we argue that these uses of *gibt* and *geht* and their corresponding forms are not related by similarity with their more basic meanings.

The largest portion of our data was annotated by the two student assistants. Between the two student assistant annotators, we report a substantial agreement ($\kappa = 0.6$), close to the agreement reported in Sanchez-Bayona and Agerri (2022, $\kappa = 0.63$). Additional posts were annotated by students of German linguistics within a seminar on metaphor. In all cases, the first author of this paper took care of the data curation, in close discussion with the annotators to resolve disagreements. In all scenarios, INCEpTION (Klie et al., 2018) was used as the main tool for annotation.

An overview of the datasets used in our experiments is given in Table 1.

4. Experimental Setup

Our main goal is to investigate methods to overcome the scarcity of data that matches our use case of posts from German religious online forums. The largest portion of this paper is dedicated to the investigation of cross-lingual transfer potentials for

Dataset	Lang.	Tokens	MRWs (%)
DE_CHR	DE	14,222	2,734 (19.22%)
EN_CHR	EN	14,981	3,170 (21.16%)
EN_VUA	EN	72,611	13,070 (18.00%)

Table 1: Overview of our data.

automatic metaphor detection within the context of genre differences. Consequently, in our experiments we train and evaluate on various combinations of English and German data. Given the lack of research on automatic metaphor detection for German, we will use the German dataset as a test set and train the models on either the smaller in-domain Reddit data or the larger, but out-of-domain VUA20 data. However, in order to gain further insight on cross-lingual transfer for this task, we also use the German data set as training set, evaluate on the English Reddit data and compare these results to the results reported in Reimann and Scheffler (2024) for training on VUA and testing on Reddit. In addition to the supervised experiments, we will also further develop previous approaches to metaphor detection with generative LLMs with in-context learning and without the use of additional training or finetuning.

Supervised Transfer Baseline. In our experiments, for the supervised and transfer learning based approach we design a simple baseline, the token-based architecture that was already used as a baseline in the 2020 Metaphor Detection Shared Task (Leong et al., 2020). It uses the contextualized BERT embeddings of a sentence, which are given to a linear layer, where a softmax predicts a label for each token (here: metaphoric or not) (BASE_XML-R). Instead of the monolingual BERT model, we use the multilingual XLM-R by Conneau et al. (2020).

Transfer Experiments. As our main experiment, we test two architectures for metaphor detection that claim to be inspired by linguistic theories and modify them slightly to enable cross-lingual transfer. On the one hand, we use MelBERT (Choi et al., 2021). It uses two transformer encoders, one to encode the entire sentence and one for the target word only, to replicate MIP and SPV. For MIP in MelBERT, the contrast between basic and contextual meaning is modeled by concatenating the contextual embedding of the target word and the embedding of the word in isolation. For SPV, Choi et al. (2021) model the contrast between a word and its context by concatenating the contextual meaning of the word and the embedding of the [CLS] token, representing the entire sentence. Both concatenations are in the end given to a linear classifier. We, however, replace the monolingual RoBERTa used in the original MelBERT with XLM-R. We prefer the original MelBERT over modifications that model the basic meaning through dictionary entries (Zhang and Liu, 2022; Babieno et al., 2022) or extract it from literal examples in the training data (Li et al., 2023). These approaches rely on either language-specific resources or training and test data in the same language and are thus not suitable for cross-lingual experiments.

As a second model, we use AdMul (Zhang and Liu, 2023). It is also influenced by MIP but approaches this in a different way. In a multi-task learning framework, they combine metaphor detection with an auxiliary task called basic sense detection, predicting if the word is used in its basic meaning. For this auxiliary task, they transform the the Sense Disambiguation (WSD) dataset of Raganato et al. (2017) into a binary dataset, where word uses with the most common sense are labeled as used in their most basic sense. The model is trained on both tasks and a global discriminator aligns the representations. We replace the DeBERTa (He et al., 2021) model of AdMul with its multilingual counterpart. For the basic meaning disambiguation task, we additionally add the German fraction of XL-WSD (Pasini et al., 2021), a multilingual extension of the WSD framework by Raganato et al. (2017). For comparison purposes and and due to a lack of separate validation data for additional hyperparameter tuning, we used the optimal hyperparameters reported in Choi et al. (2021) and (Zhang and Liu, 2023).

In all experiments, we use the available models from HuggingFace (Wolf et al., 2020). We ran all models on a NVIDIA A30 Tensor Core GPU. In total, running the experiments took approximately 90 hours: 47 hours for running the LLM experiments and 43 hours for running the experiments with the XLM-R and DeBERTa based models.

LLM Prompting. Additionally, we further develop a prompting-based LLM method for automatic metaphor detection that does not make use of any additional training data for finetuning. For this we use the method of Reimann and Scheffler (2025b) as a starting point, in particular the version that uses a one-shot prompt with an example in the last prompt. We modify their series of prompts using the insights of Hicke and Kristensen-McLachlan (2024), who observed positive results with a prompt that replicates step 3b of MIPVU, the identification of a more basic meaning. Reimann and Scheffler (2025b) modeled this step with two prompts, one to ask the model to generate a dictionary entry and another to ask which, if any, may be consid-

ered more basic according to MIPVU. We replace these two prompts with the one used by Hicke and Kristensen-McLachlan (2024), asking the model if a more basic meaning is available and, if yes, to briefly define this meaning.We provide the prompts in Appendix A. Reducing this question on a more basic meaning to one prompt would make the procedure of Reimann and Scheffler (2025b) cheaper with respect to computational costs and potentially less prone to error propagation. For all runs of our LLM approach, we use the 8B version of LLaMa 3.1, which is explicitly labeled as multilingual and among the better performing models in Reimann and Scheffler (2025a). We aim for a lightweight approach that is relatively inexpensive in computational resources. Consequently , we prefer the lightweight 8B version of LLaMa 3.1 over its 70B version.

5. Results

Table 2 shows the results for all models. First, we can see that the fine-tuned encoder-only models again outperformed the one-shot LLM, even in scenarios where training and testing data were coming from different languages. Second, the results show that cross-lingual transfer for metaphor detection is to some extent possible, with F1-scores mostly above 60%. Regarding the choices of training data, finetuning the XLM-R based models on the much smaller in-genre dataset notably outperformed finetuning on VUA for German. For evaluation on the English Reddit data, the results are more nuanced. For the simple sequence classification approach and MelBERT, the differences (between cross-genre and cross-lingual transfer) in performance are small with a slightly better precision achieved by the models trained on the smaller in-genre dataset in German. This is surprising, since the results of Reimann and Scheffler (2024) would rather suggest improvements with respect to recall. However, the best performance on the English data are achieved by AdMul with German training data and again, improvements in precision playing a major role. Overall, cross-lingual transfer (from a small in-domain dataset) shows better performance over cross-genre transfer (from a large general dataset) even for the English test set.

Finally, in terms of differences between models, we mostly observe moderate improvements over the simple XLM-R baseline by the more elaborate MelBERT and AdMul architectures, which do not fully reflect the improvement that the models achieve in monolingual settings in the reported literature. Between the multilingually adapted versions of MelBERT and AdMul we observe only minor differences in F1. However, precision and recall signal that the models indeed behave differently. When finetuned on English data and evaluated on German, AdMul appears to overgeneralize the metaphor label given the comparatively low precision, compared to the high recall values. This does not seem to be a problem for MelBERT, since it rather struggles with finding metaphorical examples in the first place. These behavioral differences also come to light when comparing the different transfer settings. For AdMul, training on the German forum dataset (DE_CHR) led to notable improvements in precision over training on VUA, while recall remains constant. However, for MelBERT, a drop in recall can be observed. In conclusion, based on the metrics, it can be stated that for finetuning on smaller datasets, AdMul appears to be the best option, where it even outperforms all models finetuned on VUA for both datasets we evaluate on.

6. Error Analysis

In order to better understand the results reported in Section 5, we conduct an extensive error analysis, considering the metaphoric words that were most commonly missed by the models (false negatives) and the literal words that were most often falsely considered to be metaphoric (false positives). For all models, these results are shown in Table 3.

We can see multiple tendencies. When finetuned on VUA, several MRWs that are characteristic for religious language are not recognized. Among the top five false negatives for MelBERT, we have both *Vater* (“father”) as a metaphor for God and its corresponding genitive form *Vaters*, echoing the results of Reimann and Scheffler (2024), as well as *Willen* (“will”), utilised in the sense of “will of God”. This is similar for AdMul, with *Vater*, as well as *Beziehung* (“relationship”), often referring to a relationship with God, and *Herrn* (“lord”) for God among the most common false negatives. Using the in-genre dataset resolves this for MelBERT and AdMul, where the number of unrecognized examples of *Vater* metaphors drops to three and zero, respectively. Interestingly, the baseline model did not profit that much from seeing the metaphorical sense of *Vater* referring to God. Smaller improvements can be seen, with 11 false negatives compared to 26 when finetuning on VUA.

In the scenario of training on the English Reddit data and testing on the German forum data, we, however, notice a sharp increase of unrecognized conventional metaphors such as *machen* (“to make” or “to do”), *klar* (“clear”) and *sagen* (“to say”). The latter may be linked to the modality of online forums. When looking up *sagen* in the German dictionary Duden, we would, on the one hand, find a sense describing the bodily action of articulating words and a more abstract sense that just describes the act of expressing or formulating something. When

Training Data	Model	DE_CHR			EN_CHR		
		P	R	F1	P	R	F1
None	BASE_LLaMa	20	63	31	22	93	36
DE_CHR	BASE_XLM-R	-	-	-	72	55	63
	MelBERT	-	-	-	68	53	59
	AdMul	-	-	-	**74**	61	**67**
EN_CHR	BASE_XLM-R	60	67	63	-	-	-
	MelBERT	**64**	62	63	-	-	-
	AdMul	56	**78**	**65**	-	-	-
EN_VUA	BASE_XLM-R	56	50	53	67	55	60
	MelBERT	56	52	54	67	54	60
	AdMul	41	72	54	57	**67**	61

Table 2: Overview of precision, recall and F1-score of the models with different combinations of training (rows) and test (columns) data.

	EN_VUA -> DE_CHR		EN_CHR -> DE_CHR		DE_CHR -> EN_CHR	
	FP	FN	FP	FN	FP	FN
Base XLM-R	da 104	Willen 26	geht 28	klar 14	father 9	thing(s) 90
	geht 33	Vater 26	lassen 12	sagen 13	see 8	way 44
	hier 26	Vaters 14	finde 12	Willen 11	start 8	feel 26
	lassen 10	Dinge 14	bleiben 11	Beziehung 11	keep 6	part 16
	Situation 9	sagen 13	wählen	Vater 11	say 6	called 15
Mel-BERT	da 108	Vater 26	geht 29	machen 14	father 10	thing(s) 83
	hier 39	Willen 26	lässt 12	sagen 13	start 9	way 40
	geht 35	Vaters 14	lassen 9	klar 11	say 8	feel 26
	gibt 15	sagen 13	Gefahrenabwehr 8	Ansatz 10	find 7	away 20
	lassen 14	Dinge 10	Pocken 7	Ansicht 9	keep 6	back 17
Ad-Mul	da 138	Vater 25	geht 25	sagen 12	father 17	feel 27
	dann 52	Beziehung 19	lassen 20	Willen 10	born 12	thing(s) 24
	gibt 39	Willen 18	Gewalt 15	Sinn 8	has 12	get 20
	hier 39	Vaters 14	bleiben 12	Eindruck 7	still 12	called 15
	mutiert 36	Herrn 8	Gebote 9	Sump 7	mother 9	elect 14

Table 3: Most frequent false positives and false negatives produced by the models across transfer scenarios.

referring to something in a forum post like in Example 1, the second sense is more suitable. However, the former sense represents a more basic meaning according to MIPVU that is also sufficiently distinct and which is arguably related by similarity, thus it would fulfill both criteria of MIPVU to be considered metaphorical.

(1) Und wenn Du nun sagst : Beleg es mir – dann kann ich nur sagen : kann ich nicht .
"And if you now say : Prove it to me – then I can only say : I can't ."

Ansatz, in the sense of "attempt" or "approach" represents an error that may be explained via cross-linguistic differences. In German, *Ansatz* may also refer to the base of something mechanical like a pipe. In the Duden, we find both this meaning and *Ansatz* referring to an attempt. As structural similarities may be drawn between the two concepts, and as the former represents a more concrete, thus basic meaning according to MIPVU, this renders *Ansatz* ("attempt, approach") to be an MRW according to MIPVU, even though it is a very conventionalized one. In English, there is no word expressing *attempt* in a metaphorical way, which would explain why models trained on English data never consider the immaterial sense of *Ansatz* to be metaphoric.

Similar observations can be made when looking at the false positives, i.e., words wrongly marked as metaphorical. For all models, using the German forum data in training and the English Reddit data in evaluation led to overgeneralization regarding family terms such as *father* and *mother*, where even literal usages of these terms were often considered metaphoric. Another striking observation regarding false positives can be made for the scenario using VUA as training data, which explains the previously observed drops in precision. Here, the adverb *da* ("there") occurs as the most frequent false negative for all models. Similar to English *there*, *da* may

occur in a spatial sense, referring to a location, as well as a temporal sense, referring to a point in time, such as the use of *da* in Example 2, or referring to situations like in Example 3. It may indeed be argued that the spatial use of *da* represents a more basic meaning according to MIPVU and that the uses of *da* referring to times and situations are related by some sort of similarity. However, none of the human annotators actually considered *da* to be metaphoric, thus raising doubts if this similarity is, from the perspective of a German native speaker, still apparent for the modern day language user.

(2) Es gibt Zeiten, da ist es egal [...].
"There are times when it does not matter [...]."

(3) Wozu brauche ich da einen eigenen Willen?
"For what do I need my own will there?"

In VUA, we can indeed find 27 instances of English *there* that were labeled as MRWs. Among these are Example 4 with *there* expressing a temporal meaning and Example 5 referring to a situation. Our results thus strongly suggest that, from these examples, the models learned that *da* in such instances may be considered metaphoric.

(4) The building society will be staffing a mortgage desk at each auction, and says buyers could arrange finance there and then, subject of course to proof of income and status.

(5) Seriously, I'd fucking have it out of there, everything I own.

Another aspect that needs to be considered is the impact of the auxiliary basic sense detection task in AdMul, which represents also the biggest architectural difference between the models. For this, we look at the five words, for which we see the biggest improvement in terms of recall when choosing AdMul over MelBERT in the in-genre cross-lingual transfer scenario and then look at the output of the basic sense detection task of AdMul for these words. For English, they are *thing*, *way*, *feel*, *away* and *back* and for German those are *machen* ("make"), *sagen* ("say"), *klar* ("clear"), *Ansatz* ("approach") and *Ansicht* ("view"). According to the linguistic theory, the metaphoric sense is never the more basic sense of a word and thus, for metaphoric words, the model should not consider them to be used in their basic meaning. Indeed, in metaphoric instances of the aforementioned words, the model predicted that they are not used in the basic sense, suggesting that it may have indeed learned valuable information for metaphor detection from the auxiliary task.

7. Discussion

From our results, we can see that architectures inspired by linguistic procedures for metaphor detection are to some extent also an option for cross-lingual metaphor detection. They also clearly outperformed our LLM-based approach, suggesting that, as long as there is training data available, such supervised approaches may be preferred. For MelBERT, we do not observe the improvements reported in Choi et al. (2021) over the simple sequence classification baseline. However, AdMul provided the strongest results overall, even though it sacrificed precision in the scenarios that involved transfer from English to German. We hypothesize that the predictions of AdMul are more closely dependent on what it saw in finetuning since our error analysis in the previous section has shown that the most frequent cases of false positives it produced were mostly the same as the Baseline and MelBERT but with a notably higher frequency.

Our results moreover underline the findings of Reimann and Scheffler (2024) that genre differences and consequently the metaphors and domain mappings represented in the training data are a major factor in metaphor detection. Since it reflects similar patterns with respect to metaphors describing God as a human, we can also see a notably better performance on the German forum data when using the small dataset from Christian subreddits for finetuning, compared to the large VUA dataset. Vice versa, finetuning on the German dataset and testing on the English Reddit data did not result in large improvements overall, which may however be expected, given its size and the fact that linguistic metaphors are still relatively language-specific. However, despite these constraints, two of the three models performed better on EN_CHR when trained on the smaller German data and the error analysis again showed patterns where the model did better on metaphors specific to religious communication.

The choice of whether to prefer cross-lingual or cross-genre transfer against a lack of perfectly suitable training data remains thus an issue depending on the specific use case. In cases where the detection of genre-specific metaphors (i.e., family terms related to God) is much more important than the detection of generally frequent, heavily conventionalized metaphors, it may be reasonable to prefer in-genre data over in-language data.

Finally, another issue raised by our results concerns the annotation process. In the example of *da*, the German annotators did not see an underlying metaphorical mapping. This also extends to the use of *there* in the English Reddit corpus, which was annotated by German speakers, as well. One hypothesis is that in both annotation projects, the

annotators were explicitly told to focus on content words and thus dismissed these examples since they resemble temporal prepositions (e.g. *at 4 pm*) expressing a TIME IS SPACE mapping very closely. On the other hand, some metaphorical mappings may be much more easy to grasp for native speakers of one language compared to native speakers of other languages. Thus, focusing on native speakers of only one language altogether would introduce unwanted biases. However, a further exploration of this question goes beyond the scope of this paper.

8. Conclusion

We collected a novel dataset of posts from Christian German online forums and annotated it for metaphor via the MIPVU procedure. We used this dataset to evaluate state-of-the-art methods for automatic metaphor detection (MelBERT and AdMul) in a cross-lingual setting. To do this, we finetuned these models on two English datasets, one smaller in-genre dataset and a larger dataset from different genres. We also used the German dataset as a training set and evaluated on the English dataset from the same genre. We compared these results to a monolingual English, cross-genre transfer setting. We also had a deeper look into our results with a comprehensive error analysis.

From these results, we conclude that cross-lingual transfer between German and English for metaphor detection is possible for metaphors where the underlying cross-domain mapping is represented in both source and target language. Moreover,in line with previous research on metaphor detection, we stress that architectures that employ contextualized embeddings and supervised learning may be preferred over zero- and few-shot approaches with LLMs as long as annotated training data is available. Regarding the specific model choice, AdMul outperformed both MelBERT and the baseline. We observed that the auxiliary task in AdMul may have helped it to learn from finetuning on smaller dataset. This is particularly relevant, given that the genre of the training data plays a vital role, to the extent that in some cases, smaller in-genre training datasets, even if they are in a different source language may be better suited than in-language data from a different domain.

For future work, we suggest, on the one hand, to explore a wider range of languages for cross-lingual transfer, especially from more dissimilar language pairings than English and German. We also suggest to consider the backgrounds of annotators more closely, especially their native language, and explore how these experiences would influence their annotations. Several other NLP tasks already benefited from including multiple perspectives instead of an aggregated gold standard (Cabitza et al., 2023) and our findings suggest that metaphor detection may be a prime example for this.

9. Limitations

One limiting aspect that prevents us from making more general statements on cross-lingual transfer and metaphor detection is the fact that we only considered two relatively related languages. Moreover, although we were aiming for a computationally cheap approach, we are aware that using larger models or additional finetuning of generative LLMs may have further improved the LLM performance here. Finally, all our annotators were native speakers of German. As we discuss in the paper, this may have brought in biases with regards to the perception of metaphors.

10. Ethical Considerations

Regarding the resource-hungry nature of generative LLMs, we limited our usage of such models to a minimum and preferred smaller models. Given that religious beliefs represent particularly sensitive personal information, we made sure that the data we used did not contain information that may link certain opinions to individual people. Our dataset also does not contain usernames. The student assistants did their annotation work within a fixed work contract and were paid according to public payscales. All annotators were informed that their annotations will be used as training data for metaphor detection.

11. Bibliographical References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions. *Applied Sciences*, 12(4).

Maria Berger, Nieke Kiwitt, and Sebastian Reimann. 2024. Applying transfer learning to German metaphor prediction. In *Proceedings of the 2024 Joint International Conference on Computational*

Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1383–1392, Torino, Italia. ELRA and ICCL.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5850–5860, Miami, Florida, USA. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

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.

Markus Egg and Valia Kordoni. 2023. A corpus of metaphors as register markers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 220–226, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

J. Berenike Herrmann, Karola Woll, and Aletta G. Dorst. 2019. Chapter 6. Linguistic metaphor identification in German. In Susan Nacey, Aletta G. Dorst, Tina Krennmayr, and W. Gudrun Reijnierse, editors, *MIPVU around the world*, pages 113–135. John Benjamins Publishing Company.

Rebecca M. M. Hicke and Ross Deans Kristensen-McLachlan. 2024. Science is exploration: Computational frontiers for conceptual metaphor theory. In *Proceedings of the Computational Humanities Research Conference 2024*.

Anna Hülsing and Sabine Schulte Im Walde. 2024. Cross-lingual metaphor detection for low-resource languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Matej Klemen and Marko Robnik-Šikonja. 2023. Neural metaphor detection for slovene. In *Selected papers from the CLARIN Annual Conference 2022*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.

Volkhard Krech, Tim Karis, and Frederik Elwert. 2023. Metaphors of religion. a conceptual framework. *Metaphor Papers*, 1.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Univ. of Chicago Press, Chicago [u.a.].

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Jiahui Liang, Aletta G. Dorst, Jelena Prokic, and Stephan Raaijmakers. 2025. Using gpt-4 for conventional metaphor detection in english news texts. *Computational Linguistics in the Netherlands Journal*, 14:307–341.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2024. Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.

Sebastian Reimann and Tatjana Scheffler. 2025a. The struggles of large language models with zero- and few-shot (extended) metaphor detection. *Journal for Language Technology and Computational Linguistics*, 38(2):97–109.

Sebastian Reimann and Tatjana Scheffler. 2025b. Using large language models to perform MIPVU-inspired automatic metaphor detection. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert. In *Text, Speech, and Dialogue*, pages 104–111, Cham. Springer International Publishing.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor

detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.

A. Prompts

Table 4 gives an overview over the prompts used for metaphor detection on the English data and Table 5 shows the prompts used for metaphor detection on the German data, which are the translated equivalent of the English prompt but with a different example for the final prompt (***Weg*** (“way") instead of *journey*).

Step	Prompt
MIPVU Step 3a	In one sentence, describe the meaning of the given word in the context of the given post as general as possible. Word: [WORD] Post: [POST]
MIPVU Step 3b	For the given word, determine if it has a more basic contemporary meaning in other contexts than the one in the given post. For our purposes, basic meanings tend to be: - more concrete; what they evoke is easier to imagine, see, here, feel, smell, and taste - related to bodily action - more precise (as opposed to vague) - historically older Basic meanings are not necessarily the most frequent meanings of the word. If such a more basic meaning can be identified, output a brief definition of this basic meaning, otherwise just output 'No.' Word: [WORD] Post: [POST]
MIPVU Step 3c	Can you see a similarity between the senses 1 and 2? 'Similarity' may also mean that the two senses denote distinct concepts that share certain aspects, functions or features. The following example for the word 'journey' illustrates this: journey: Sense 1: "an occasion when you travel from one place to another, especially over a long distance" Sense 2: "a long and often difficult process by which someone or something changes and develops" Answer: Yes. Sense 1 and Sense 2 are similar because in both senses refer to something that takes a longer period of time. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONTEXTUAL SENSE] Sense 2: [MORE BASIC SENSE]

Table 4: Overview over the English prompts used.

Step	Prompt
MIPVU Step 3a	Beschreibe in einem Satz die Bedeutung des gegebenen Wortes im Kontext des gegebenen Posts. Bleibe dabei so generell wie möglich. Wort: [WORD] Post: [POST]
MIPVU Step 3b	Entscheide für das gegebene Wort, ob es eine grundlegendere, moderne Bedeutung in anderen Kontexten besitzt, als die im gegebenen Post. Für unsere Zwecke ist eine grundlegendere Bedeutung eine Bedeutung, die: - konkreter ist; das heißt, das was sie evoziert kann man sich leichter vorstellen und kann man sehen, hören, riechen, fühlen oder schmecken - die mit körperlichen Handlungen verbunden ist - die präziser ist - die älter ist Eine grundlegendere Bedeutung muss nicht zwangsweise die häufigste Bedeutung des Wortes sein. Falls so eine Bedeutung identifiziert werden kann, dann definiere diese kurz. Andernfalls, falls dem nicht so ist, antworte mit "Nein.". Wort:[WORD] Post:[POST]
MIPVU Step 3c	"Gibt es eine Ähnlichkeit zwischen den beiden Bedeutungen 1 und 2? "Ähnlichkeit" in diesem Zusammenhang, meint auch, dass die beiden Bedeutungen sich auf unterschiedliche Konzepte beziehen, aber bestimmte Aspekte, Funktionen oder Merkmale teilen. Das folgende Beispiel für das Wort "Weg" beschreibt dies genauer: Weg: Bedeutung 1: "Strecke, die zurückzulegen ist, um an ein bestimmtes Ziel zu kommen" Bedeutung 2: "Art und Weise, in der jemand vorgeht, um ein bestimmtes Ziel zu erreichen; Möglichkeit, Methode zur Lösung von etwas" Antwort: Ja, die Bedeutungen 1 und 2 sind sich ähnlich, da Bedeutungen ein Ziel am Ende steht. Antworte mit "ja" oder "nein", gefolgt von einer kurzen Erklärung. Bedeutung 1: [CONTEXTUAL SENSE] Bedeutung 2: [MORE BASIC SENSE]

Table 5: Overview over the German prompts.

Decomposing Creativity: Two Small Datasets Combining Originality Ratings and Metaphor Annotations

Emilie Sitter, Sina Zarrieß, Omar Momen, Berenike Herrmann
CRC 1646 – Linguistic Creativity in Communication
Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{emilie.sitter,sina.zarriess,omar.hassan,berenike.herrmann}@uni-bielefeld.de

Abstract

We introduce MetaphOrig, two small datasets comprising two genre-specific collections of spatial descriptions for the study of linguistic creativity and Non-Literal Expressions (NLEs). The sentence-level spatial descriptions were extracted from two distinct genre- and time-specific source corpora. Both source corpora comprise German texts: literary prose from the 18th to the 20th century (KOLIMO) and factual travel reports from the 21st century (Wikivoyage). Along with the spatial descriptions, the datasets contain sentence-level originality ratings obtained through crowdsourcing and from four different LLMs (GPT-5, Qwen2.5-32B-Instruct, Mistral-Small-3.2-24B-Instruct, and Llama-3.2-3B), and word-level metaphor annotations. We provide the MetaphOrig datasets, including all annotations, to the community. The datasets can be used for further research on linguistic creativity or metaphor, either in one specific textual domain or comparatively across the two domains. We conduct an illustrative study on the datasets, treating originality as a proxy of textual creativity. In both datasets, we investigate potential correlations between sentence-level originality ratings and the density of metaphorical expressions within each sentence. We find the correlation to be present only in the KOLIMO dataset. A comparison of human and LLM originality ratings shows that this pattern holds for both types of ratings.

Keywords: Linguistic Creativity, Annotation, Crowdsourcing, Metaphor, Originality, MIPVU, Large Language Models

1. Introduction

This paper introduces MetaphOrig, two small datasets that aim to bridge the gap between holistic, crowdsourced originality ratings at the text-level and a fine-grained, linguistic analysis of individual rhetorical devices. The exploratory study emphasizes metaphors, which are among the most prominent Non-Literal Expressions (NLEs) as a potential contributor to linguistic creativity.

According to the standard definition of creativity, the phenomenon comprises the two dimensions of originality and effectiveness (Runco and Jaeger, 2012). In this paper, we particularly focus on originality as its presumably most important dimension (Diedrich et al., 2015). We understand originality as the perceived novelty, unconventionality, or unexpectedness of a text for a specific reader.

Studies in psychology or psycholinguistics often aim to assess the creativity of ideas, artistic works, or linguistic productions. To this end, human assessments of creativity are oftentimes collected through crowdsourcing or by ratings of domain experts (Qian and Plucker, 2017).

In contrast, linguistics and NLP rarely conceptualize creativity as a holistic property of an idea or utterance. Instead, they tend to focus on specific textual features and formal characteristics that may influence how creative an utterance is perceived to be by humans (Weinstein et al., 2022; Zedelius et al., 2019). Metaphors provide a particularly salient example of such a feature, as they are often assumed to enhance the perceived creativity of a sentence. They are one of the most extensively researched forms of potentially creative and original language use (Kohl et al., 2020), and they are widespread or even ubiquitous in language and thought (Lakoff and Johnson, 2003). Yet, in NLP research, there is relatively little work and hardly any existing datasets that combine annotations of metaphors, on the one hand, and of creativity or originality, on the other hand.

According to Lakoff and Johnson's Conceptual Metaphor Theory (CMT), metaphor is a cognitive mechanism that structures human understanding by mapping complex experiences onto more concrete or familiar domains. Metaphorical expressions vary widely across genres in form, function, and cognitive effects, as well as in their degree of creativity or originality (Steen et al., 2010a; Herrmann, 2015; Momen et al., 2026). Even in presumably creative domains such as advertisement or literary texts, many metaphorical expressions are conventional, but still contribute to perceived creativity (Steen et al., 2010a; Dorst, 2015; Burgers et al., 2015). In the context of NLP, metaphors thus continue to pose challenges for current approaches to annotation and detection because of their context dependence and varying degrees of conventionality (Maudslay and Teufel, 2022; Ye et al., 2025).

The MetaphOrig datasets consist of two small, genre-specific datasets of sentence-level descriptions of spatial scenes, annotated with originality ratings and metaphor labels. Spatial descriptions

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 119–132
11 May 2026.

Rating Item	Example 1 (Literary)	Example 2 (Non-literary)
Original Text	*Die Glastüren zur Veranda standen offen, und der Duft des Flieders drang herein, der wie eine Mauer aus weißem und hellblauem Gewölk den Garten einhegte.*	*São Roque: Diese Stadt liegt an der Nordküste von Pico und ist bekannt für ihre schönen Naturschwimmbecken, in denen Besucher baden können.*
Translation (own)	*The glass doors to the terrace were open, and the scent of the lilac drifted in, enclosing the garden like a wall of white and light blue clouds.*	*São Roque: This town is located on the north coast of Pico and is famous for its beautiful natural pools where visitors can swim.*
Source	Eduard von Keyserling, *Abendliche Häuser* (1940)	Wikivoyage, *Pico* (2023)
Metaphors	Metaphor Density: 0.23913 Direct Metaphors: 4 ("Mauer", "weißen", "hellblauen", "Gewölk") Indirect Metaphors: 1 ("standen offen")	Metaphor Density: 0.047619 Direct Metaphors: 0 Indirect Metaphors: 1 ("liegt")
Originality Ratings	Humans: 4.8 GPT-5: 5.0 Qwen2.5-32B-Instruct: 4.6 Mistral-Small-3.2-24B-Instruct: 5.0 Llama-3.2-3B: 4.6	Humans: 2.1 GPT-5: 1.0 Qwen2.5-32B-Instruct: 2.0 Mistral-Small-3.2-24B-Instruct: 3.0 Llama-3.2-3B: 3.3

Table 1: Example sentences from the MetaphOrig datasets with respect to metaphors and averaged originality ratings by humans and LLMs.

commonly appear as textual elements in different types of genres, including travel reports or novels. Such descriptions are comparable content-wise and can be understood in isolation. We therefore assume that human raters can assess the linguistic originality of these controlled sentences more easily than that of randomly selected passages which would differ substantially in terms of content.

The spatial descriptions were extracted from two fundamentally different textual domains, yielding two genre-specific datasets: 18th to 20th century literary prose on the one hand and factual travel reports on the other (see the example sentence in Table 1 from a literary prose text alongside a non-literary sentence from a travel report). The older literary prose presumably seems highly literary to contemporary readers, while the travel reports are characterized by factual and, presumably, more non-figurative language.

For each sentence in both datasets, we additionally collected originality ratings via online crowdsourcing and annotated word-level metaphors following the MIPVU procedure (Steen et al., 2010b; Herrmann et al., 2019). These annotations result in two original datasets of sentences matched with both sentence-level originality scores and word-level metaphor information.

We showcase a potential use-case of the MetaphOrig datasets by presenting an analysis that correlates metaphor density with originality according to the human judgments as well as to LLM ratings of originality. In a first step, the two very different datasets are used to analyze the two formally distinct textual domains in terms of originality and metaphoricity. We examine whether the relationship between human originality ratings and metaphor use varies across them. In a second step, we explore whether LLMs can be leveraged for rating originality similarly to a crowdsourced collection of ratings. In addition to the human originality ratings, we thus collect ratings from four different LLMs. We aim to explore not only the extent to which LLMs approximate human originality judgments, but also whether their originality ratings appear to be grounded in similar textual cues.

2. Background

2.1. Creativity Assessment

In creativity research, the goal of creativity assessments is to determine the creativity of particular products, as opposed to, e.g., the creativity of individuals (such as writers) or processes (their writing processes). The creativity of products, such as texts, is typically assessed through human ratings. Most of these human-evaluation approaches include Likert scale ratings of novelty or originality collected from experts or via crowdsourcing (Hennessey et al., 2011; Kaufman et al., 2009), structured pairwise comparisons where raters select the more creative item from a pair (Crompvoets, 2025; Cao et al., 2026), and ranking tasks in which multiple items are ordered according to perceived creativity or originality (Do Dinh et al., 2018). One of the most influential approaches in this area is the Consensual Assessment Technique (CAT) introduced by Amabile (1982). While CAT raters are usually domain experts, a certain degree of subjectivity in the ratings is inevitable (Qian and Plucker,

2017). Mozaffari (2013) proposed an analytical rubric that assesses multiple sub-dimensions of creativity rather than relying on one single, holistic creativity score. All these approaches share the common assumption that creativity can be inferred from the collective judgments of human raters.

2.2. Metaphors and Creativity

The most influential approach to metaphor in linguistics is the Conceptual Metaphor Theory (CMT), introduced by Lakoff and Johnson in their seminal work *Metaphors We Live By* (1980) (Lakoff and Johnson, 2003). In linguistics, metaphor is still predominantly understood according to CMT as a cognitive mechanism by which a (typically abstract) *target domain* is conceptualized in terms of a more concrete *source domain*. While conceptual metaphors, the underlying cognitive mechanisms of a metaphor, are typically highly conventional, their linguistic realizations can vary widely and may be realized as novel and creative utterances (Kövecses, 2020). This can be observed especially in literary text which allows authors and poets to create novel linguistic metaphors by creatively re-interpreting conventional and everyday conceptual metaphors (Semino and Steen, 2008; Kövecses, 2010).

The Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) and its extension by VU Amsterdam (MIPVU) (Steen et al., 2010b) are currently among the most common ways to annotate metaphor in linguistics. However, these annotation procedures do not capture the degree of creativity, novelty, or deliberateness of a metaphor. Existing approaches to annotate metaphor novelty as one dimension of creativity are based either on crowdsourcing (Do Dinh et al., 2018) or they are dictionary-based (Egg and Kordoni, 2022). According to Reimann and Scheffler (2024), dictionary-based annotations align more reliably with expert metaphor annotations.

Further, more detailed annotation protocols of metaphors are based on Deliberate Metaphor Theory (DMT) (Steen, 2017) and aim to identify whether a metaphor is used potentially deliberately (Reijnierse et al., 2018; Dipper et al., 2024). DMT suggests that non-deliberate metaphors are always conventional.

In the context of working with potentially creative or deliberate metaphors, the distinction between direct and indirect metaphors introduced by MIPVU is also worth considering (Steen et al., 2010a). Direct metaphors, often comprising longer phrases, are typically realized in the form of similes (comparisons) and explicitly marked in the discourse, e.g., by 'as' or 'like' (Steen et al., 2010a) (see example (1) in Table 1). They are particularly important when studying creativity because they might be strongly related to originality and deliberate usage. Indirect metaphors, by contrast, largely occur on the word and phrase level and are far more frequent, conventionalized, and more often used non-deliberately (Steen, 2017). They are less likely to contribute strongly to a sentence's overall originality.

When creating the MetaphOrig datasets, we opted for a parallel annotation setting: we collected standard creativity ratings on the sentence-level, and independent metaphor annotations on the word level according to MIPVU.

2.3. LLMs in creativity and metaphor research

To date, a variety of computational approaches have been developed to assess creativity. Acar (2025) provides an overview of the evolution of computational methods for assessing creativity.

Research on using LLMs to predict creativity ratings has so far come mainly from the social sciences and from psychology. Many of these experiments rely on human creativity ratings of either texts of different domains or of responses to creativity tests, such as the Alternative Uses Task (AUT). Luchini et al. (2025) demonstrated that LLMs outperform measures of semantic distance in terms of predicting human originality ratings of creativity test responses. Organisciak et al. (2023) found similar results for AUT responses and Laverghetta et al. (2025) for solutions to design problems. Rabeyah et al. (2024) point out the high correlation of multiple LLMs with each other in creativity scoring of AUT responses. More specifically, in the humanities, in a study with a particular focus on creative writing, Kim and Oh (2025) could demonstrate high consistency and performance of LLM ratings. Not specifically related to originality, but in terms of assessing the quality of narrative text, Chiang and Lee (2023) demonstrated that LLMs are a potential alternative to human ratings.

Specific work on annotating metaphors or their originality or novelty in the context of NLP and Digital Humanities has been carried out by several researchers. Some studies aimed to apply LLMs as annotators of word-level metaphors (Hicke and Kristensen-McLachlan, 2024; Reimann and Scheffler, 2025; Sánchez-Montero et al., 2025). They demonstrate that under certain conditions, LLM-based annotations can approximate human ones, but also highlight the brittleness of prompting approaches and the sensitivity to specific prompts. DiStefano et al. (2024) fine-tuned two language models on human creativity ratings of metaphors, suggesting that these models may be able to approximate certain aspects of how humans interpret figurative language. Momen et al. (2026) explored LLM surprisal as a predictor of metaphor novelty. They found a positive correlation between word-

level surprisal of metaphors and their novelty ratings in multiple datasets. In the present paper, we focus on comparing LLM originality ratings to human ratings, leaving the modeling of (creative) metaphor detection for future work.

3. Introducing the MetaphOrig datasets

This section describes the construction of the two datasets of MetaphOrig. We provide the datasets, including all annotations on GitHub.

3.1. Spatial descriptions datasets

Both MetaphOrig datasets consist of 100 spatial descriptions extracted from German texts. They were selected randomly from two larger datasets of sentences that have been annotated manually based on annotation guidelines for identifying descriptions of spatial scenes (Sitter et al., 2025). Such spatial descriptions aim to describe concrete, static spatial, scenic surroundings without any immediate actions. We focus on spatial descriptions for several methodological reasons: First, spatial descriptions are common and integral textual elements across different genres and can be found in both source corpora of the MetaphOrig datasets. Second, their comparability in terms of content facilitates the assessment of originality at the linguistic level, as it reduces the risk that raters conflate linguistic creativity with originality of the content. Third, the selected spatial descriptions are comprehensible in isolation and can be presented to raters without additional textual context. This restriction allows for stronger experimental control and ensures a substantial degree of comparability between individual items, although it may limit the generalizability of the findings to entire texts.

3.2. The KOLIMO dataset

The first small dataset consists of 100 spatial descriptions that originally were extracted from KOLIMO, the "Corpus of Literary Modernity", (Horstmann, 2019; Horstmann and Akazawa, 2024; Herrmann and Lauer, 2018). This corpus comprises literary fictional prose texts mainly from the 18th to 20th centuries, while most of the sentences have been extracted from texts published in the 19th century. We chose a literary source corpus for building the dataset because metaphors are a particularly important rhetorical device in literary writing and we expected a high amount of non-literal expressions in these texts. Moreover, to the best of our knowledge, there is currently no prior study that systematically collected originality ratings for individual literary sentences using crowdsourcing methods. Using 18th–20th century language complicates comparisons with contemporary texts. However, it is difficult to obtain and distribute annotated data for contemporary literary texts due to copyright reasons. This limitation particularly affects highbrow canonical works, which are often regarded as the most creative representatives of their genre. Consequently, we opted for texts that are both publicly accessible and literary prestigious in character instead of contemporary text material.

3.3. The Wikivoyage dataset

The second small dataset consists of 100 spatial descriptions extracted from the Wikivoyage corpus (Nolda, 2024; Wikimedia Foundation Inc., 2025). Wikivoyage is a collection of non-literary travel reports (21st century). Most sentences are factual and plain (see example (2) in Table 1). This dataset can thus be expected to contain less figurative and less original language use, compared to the KOLIMO dataset.

4. Ratings and annotations

For each of the textscMetaphOrig datasets, we collected originality ratings via crowdsourcing and from four different LLMs. The metaphors in the sentences were annotated by trained experts following the MIPVU procedure (Steen et al., 2010b). This section describes the obtained ratings and the annotations.

4.1. Human originality ratings

We collected the ratings for all sentences in both small datasets as part of a large-scale rating study on linguistic creativity on Prolific. Prolific is a crowdsourcing platform that is primarily used in the social sciences and aims to ensure the high quality of the data collected through quality checks. Remuneration for the raters was calculated based on the German minimum wage at the time of data collection (May 2025).

A total of 120 L1 German speakers rated all items on a six-point Likert scale for originality. 10 people in total rated each individual item in an incomplete between-groups design that was applied to reduce the cognitive load on the participants.

Participants were presented with 50 items each (one third each from the KOLIMO and the Wikivoyage dataset, one third filler items). We provided no explicit genre information and stated that all presented items were taken from everyday language texts. This was intended to ensure a uniform extra-textual context for both datasets and thus maximize the comparability of the ratings. At the same time,

our aim was to examine the extent to which textual genre cues alone affect originality judgments, without explicitly priming raters toward a literary reading mode or evoking genre expectations that might influence their ratings (Knoop and Blohm, 2025). However, given that humans are often able to infer genre solely by reading a text without relying on any explicit extra-textual genre cues (Knoop et al., 2024), the KOLIMO example (1) in Table 1 is still likely to be considered as literary but not the Wikivoyage example (2). Some participants explicitly commented in the feedback field at the end of the study that they identified the KOLIMO sentences as passages from literary text despite being presented as everyday language.

In addition to a German instruction text (Appendix A), participants received examples of one very original and one very conventional sentence. Rating each individual sentence was not mandatory. The mean number of raters per sentence is thus 9.97 (range 9–10, SD = 0.171). The inter-rater agreement is moderate (Krippendorff's Alpha = 0.525), likely reflecting the inherently subjective nature of perceived originality.

Importantly, we do not aim to model how texts of the KOLIMO dataset were perceived at the time of their publication. We are rather interested in modeling the "synchronous" originality judgments of today's readers which are made from a contemporary perspective. The temporal distance to the texts' publication presumably contributes to a higher degree of literariness as perceived by today's readers. From a contemporary perspective, literariness may not always be clearly separable from originality. We assume that the perceived literariness of these historically distant texts may contribute to higher originality ratings among today's readers. By contrast, readers at the time of publication may not necessarily have perceived texts with similar stylistic features as equally marked or original.

4.2. LLM originality ratings

We obtained LLM annotations by three instruction-tuned locally hosted LLMs, covering a spectrum of different sizes, and one commercial LLM, GPT-5. The largest locally hosted model, Qwen2.5-32B-Instruct (Yang et al., 2024), has previously proven particularly well suited for the automatic extraction of spatial descriptions (Sitter et al., 2025). This suggests that it might be well adapted to this kind of data. Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025) represented another family of LLMs and a medium-sized model. The smallest model, Llama-3.2-3B (et al., 2024), was chosen to investigate how well a relatively small LLM works for this task.

All LLM prompts were standardized as much as possible to emulate the human annotation process. The models received basic information about the rating study in the *system prompt*. The *user prompt* contained the same instructions the human raters received. Following a few-shot prompting approach, the user prompt contained examples for one highly original and one highly conventional sentence. In addition to the instructions for human raters, we specified the desired output format in the LLM instructions (see appendix B).

To approximate the Prolific study setup with 10 human raters per sentence, we prompted each model 10 times per sentence. Temperature was set to 1.0 to introduce variability and approximate the diversity of human raters. A final score per sentence for each model was obtained by averaging all 10 ratings.

4.3. Metaphor annotations

We annotated metaphors at the word level applying the Metaphor Identification Procedure of VU Amsterdam (MIPVU) (Steen et al., 2010b) and particularly its specification for German (Herrmann et al., 2019). Applying MIPVU, annotators assessed for each individual word in a sentence whether its contextual sense shows a meaning that is distinct from its more basic, typically concrete sense, while understood by comparison to it (such as "*enclosing*" in example (1), referring to a scent instead of a physical enclosure). More basic senses were looked up in Duden.

All metaphor information was annotated collaboratively by three different annotators trained on the MIPVU procedure (one of whom was the first author of this paper). The annotators cross-checked each other's annotations; cases of disagreement were discussed in detail.

MIPVU allows to annotate both direct and indirect metaphors regardless of their degree of conventionality or novelty. Dictionary-based approaches to annotating metaphor novelty would not be able to capture all direct metaphors and possibly fail to do justice to the creative potential of literary texts. The main annotation categories are "Non-MRW" for words used in their literal sense, "indirect", and "direct", as well as "WIDLII" ("When in doubt, leave it in.") for borderline cases. Markers of direct metaphors are annotated as "Mflag". Following MIPVU, each word of a direct metaphor has to be annotated as metaphorical.

MIPVU annotations can be used to calculate a "metaphor density" for each sentence. Metaphor density is the ratio of metaphorically to non-metaphorically used words in a sentence. To prevent the metaphor density score from being overly skewed toward direct metaphors, we consider only content words and annotated non-content words within direct metaphors as "stopwords". Stopwords within direct metaphors can themselves be indirect metaphors and therefore annotated as "indirect":

	KOLIMO	Wikivoyage
Mean sentence length	23.38	16.15
Median sentence length	21.00	16.00
Metaphor words overall	16.60%	8.57%
Direct metaphor words	2.46%	0.00%
Indirect metaphor words	14.14%	8.57%
Mean metaphor density per sentence	0.1374	0.0870
Sentences with at least one metaphorical word	86%	69%

Table 2: Descriptive statistics for datasets of MetaphOrig

(1) In der Moderluft schlotternd, sahen sie Fröschlach als eine Stadt aus grünen Kachelöfen [...].
Shivering in the cold air, they saw Fröschlach as a city of green tiled stoves [...]. (own translation)

In Example (1), the word *als* 'as' is annotated as "Mflag", marking the onset of a direct metaphor. The phrase *eine Stadt aus grünen Kachelöfen* 'a city of green tiled stoves' constitutes the direct metaphor as a whole, whereas the article *eine* 'a' and the preposition *aus* 'of' can be treated as stopwords. Strictly applying MIPVU, the preposition *aus* itself is a (highly conventional) metaphor and thus annotated as "indirect".

When calculating the metaphor density, we do not differentiate between direct and indirect metaphors. Ambiguous metaphors ("WIDLII") are counted as half. Multi-word expressions are treated as single words. The metaphor density thus represents a comparable score of metaphoricity for all sentences across both entire datasets (Steen et al., 2010a).

5. Analysis

To demonstrate a potential use of MetaphOrig, we analyze the relationship between human and LLMs' originality ratings and examine the role of metaphors in shaping these ratings. All statistical analyses were conducted in R. Relationships between metaphor density, human, and LLM originality ratings were modeled using Cumulative Link Models (CLMs), which allow the analysis of ordinal data. We used the `ordinal` package for ordinal regression models (Christensen, 2023). To facilitate comparison across models, we also compute Spearman's rank correlations. To ensure comparability across corpora and raters, metaphor density and all ratings were rescaled to the same range using the `rescale()` function from the `scales` package in R (Wickham et al., 2011).

5.1. Results

Descriptive Statistics The source and genre differences between the datasets of MetaphOrig yield substantial differences in linguistic form. Even very basic descriptive statistics reflect that the spatial descriptions extracted from KOLIMO are highly literary, i.e., descriptions from KOLIMO exhibit a higher average sentence length and a higher degree of metaphoricity (Table 2) in comparison to Wikivoyage descriptions. Yet, it is worth noting that even in Wikivoyage, 69% of sentences feature at least one metaphor.

Previous comparative research has shown differences in metaphor type use between genres. Even though literary texts do not necessarily exhibit the highest overall metaphor density, they do have the highest frequency of direct metaphors (Steen et al., 2010a; Dorst, 2015; Herrmann, 2015). Such a difference in the use of direct metaphors is also evident in our datasets (Table 2). Since most direct metaphors are deliberate, literary spatial descriptions such as those from KOLIMO can also be expected to contain considerably more deliberate metaphors.

Figure 1 reports the mean, median, range, and standard deviation of the averaged human and LLM originality ratings for each small dataset. Mean and median ratings in each rating setting are higher for the KOLIMO dataset than for the Wikivoyage dataset.

The boxplots also show bigger ranges of originality ratings (by both humans and LLMs) for the KOLIMO dataset than for the Wikivoyage one. Moreover, they reveal that only humans and GPT-5 assign ratings of 1 (not original at all), while the other three LLMs assign ratings of at least 1.6. LLMs tend to give higher ratings than humans, except GPT-5 in the Wikivoyage dataset.

Correlating human originality ratings and metaphor density To analyse the effect of word-level metaphors on human originality ratings of full sentences, we modeled humans' and LLMs' originality ratings with CLMs as an ordinal outcome predicted by the metaphor density, the dataset, and these predictors' interaction. Table 3 reports the effect of metaphor density on the originality ratings in both datasets. Figure 2 displays the human originality ratings by metaphor density, along with the ordinal regression lines computed by the CLM. Similar plots for all rater LLMs can be found in Appendix C. Only in the KOLIMO dataset, metaphor density has a significant positive effect on the human originality ratings. The Wikivoyage dataset shows no effect of metaphor density on originality ratings. While the human raters consistently rate KOLIMO sentences as more original than Wikivoyage sentences even when not metaphorical at all, Figure 2 shows that as

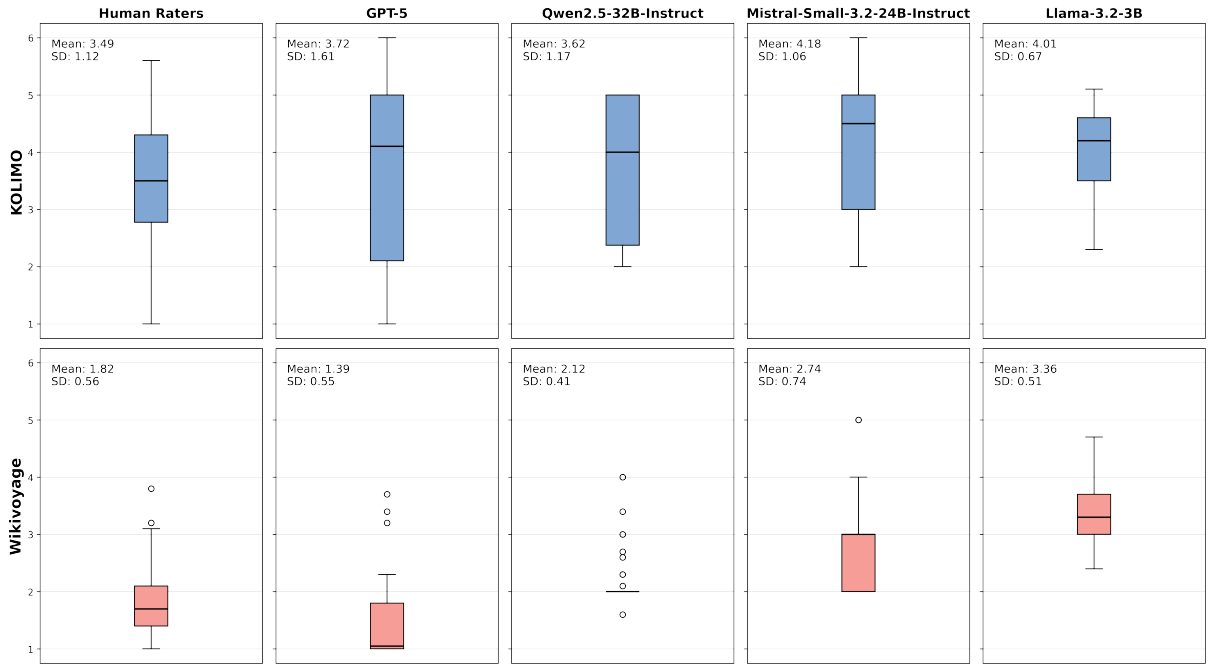


Figure 1: Summary statistics for originality ratings (scale 1–6) for humans and LLMs across the averaged ratings.

Rater	Data	β-coef	SE
Human Ratings	KOLIMO	5.57***	1.00
	Wikivoyage	0.79	0.94
GPT-5	KOLIMO	4.60***	0.89
	Wikivoyage	1.10	0.96
Qwen2.5-32B	KOLIMO	4.21***	0.99
	Wikivoyage	0.83	1.54
Mistral-Small-3.2-24B	KOLIMO	3.94***	0.99
	Wikivoyage	-0.51	0.96
Llama-3.2-3B	KOLIMO	2.06**	0.78
	Wikivoyage	-0.26	0.93

Table 3: Effect of word-level **metaphor density** on originality ratings (from humans and instruct-tuned LLMs). Higher β-coefficients indicate that higher metaphor density predicts higher originality ratings. Stars denote significance (*$p < .05$, **$p < .01$, ***$p < .001$). Standard Error (SE) quantifies the uncertainty of the estimated coefficient.

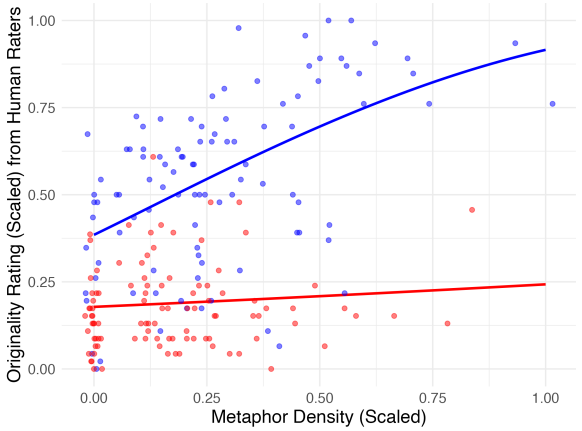


Figure 2: Averaged human ratings per sentence and ordinal regression lines for each corpus (blue = KOLIMO, red = Wikivoyage)

metaphor density increases, the gap between the KOLIMO and Wikivoyage dataset regression lines widens. This illustrates that (i) literary language is generally perceived as more original by humans in our data, and that (ii) within the literary genre, there is a strong impact of metaphors on originality for human readers.

Correlating human and LLM originality ratings To assess the alignment of human and LLM originality ratings, we fitted one CLM for each rater LLM with the human rating as ordinal dependent variable and the LLM rating, the source corpus (KOLIMO vs. Wikivoyage), and their interaction as predictors. These complex models provided significantly better fits than simpler alternatives with less predictors. We also computed Spearman rank-order correlations between each LLM's ratings and human ratings. Results of the CLMs and Spearman correlations are reported in Table 4. The beta coefficients of the CLMs indicate significant alignment between humans and LLMs for both datasets. Rank-order correlations between human and LLM ratings are stronger for the KOLIMO dataset across all four models. GPT-5 shows the highest correlation with human ratings in the KOLIMO dataset, followed by

Qwen2.5-32B-Instruct, whereas Mistral-Small-3.2-24B-Instruct shows the highest correlation in the Wikivoyage one.

Correlating LLM originality ratings and metaphor density The analysis shows that metaphor density has a generally positive effect on the ratings; i.e. the higher the metaphor density, the higher the originality rating. Just as for the human raters, this effect is significant only in the KOLIMO dataset (highly significant for GPT-5, Qwen2.5-32B-Instruct, and Mistral-Small-3.2-24B-Instruct; significant for Llama-3.2-3B) but not in the Wikivoyage dataset. Similar to humans, LLMs rate sentences in the KOLIMO dataset as more original, regardless of their metaphor density—however, the higher the metaphor density, the stronger its impact on originality ratings (see Figure C).

5.2. Discussion

Using the MetaphOrig datasets, the showcase study investigated correlations between originality ratings and metaphor density, as well as between human ratings and LLM ratings.

We find that in general, sentences in the literary KOLIMO dataset are perceived as considerably more original (Figure 1). The statistical analysis also gives reason to assume that originality correlates with metaphor density in the literary KOLIMO dataset. This effect does not hold in the Wikivoyage dataset (Table 3). Generally, these results indicate interesting and potentially complex relationships between creativity, NLEs, and genre, calling for more research on this understudied topic.

Even though the MIPVU annotations do not provide information on the deliberateness or novelty of metaphors, the absence of direct metaphors in the Wikivoyage dataset and the presence of similes in the KOLIMO dataset (Table 2) may partially explain the observed differences. Because of their textual literariness (comparably more frequent and more patterned rhetorical devices), we assume that texts in the KOLIMO source corpus generally contain more deliberate metaphors. A promising direction for future research would be to explicitly annotate and systematically analyze this type of metaphor in the MetaphOrig datasets.

Correlating human and LLM ratings of sentence originality demonstrates that LLMs' assessments significantly align with how humans rate originality. This alignment is much stronger in the KOLIMO dataset than in the Wikivoyage travel report dataset. Just as for human raters, higher LLM originality ratings in the KOLIMO dataset appear to be associated with greater metaphor density. This aligns with our aim of investigating whether LLMs rely on textual cues similar to those underlying human originality judgments, and suggests that metaphor density may function as one such shared cue. Among the LLMs tested, GPT-5 shows the highest Spearman correlation with human ratings for the KOLIMO dataset. For the non-literary Wikivoyage dataset, Mistral-Small-3.2-24B-Instruct-2506 performs best, however, with GPT-5 yielding almost similar results. Overall, this suggests that for this specific rating task of sentence originality, GPT-5 is the most suitable model (Table 4). However, since GPT-5 can be accessed only via OpenAI's API, its usage incurs a certain monetary cost ($4.48 for the batch API in total for both datasets).

One possible further usage of the introduced small datasets could be a more explicit focus on the application of LLMs for metaphor identification in different textual domains. Using an LLM to generate the metaphor annotations that we obtained manually would enable a direct comparison between human and model-based annotations.

Importantly, in this illustrative study, we do not aim to explain any variation in both human and LLM originality ratings by metaphors alone. Originality is likely influenced by a range of interacting stylistic and rhetorical features, such as sentence length, grammatical patterns, or repetition schemas that were not controlled in this study. While the results indicate that metaphors are a particularly salient feature for perceived originality in literary texts, other linguistic features may play a more prominent role in non-literary domains. Future work with the introduced small datasets should therefore consider metaphor as only one feature of a broader set of rhetorical and stylistic devices that can occur in the sentences.

In addition, many sentences in the KOLIMO dataset might automatically seem more poetic and thus more original because of their temporal distance from contemporary language. While identical analyses can be applied to both introduced datasets, this aspect of their design should therefore be treated with caution. In future work, the dataset collection could be expanded with additional genre-specific subsets. This would enable more fine-grained comparisons between different types of literary genres (e.g., highbrow and lowbrow literature) and allow for more systematic analyses across historical periods.

6. Conclusion

We introduce the MetaphOrig datasets, consisting of two small datasets of sentence-level spatial descriptions from different source corpora. We provide the sentences, sentence-level originality ratings from humans and LLMs, and word-level metaphor annotations following the MIPVU procedure.

Model	Corpus	β-coefficient	SE	Spearman Correlation
GPT-5	KOLIMO	8.76***	0.90	**0.8306**
	Wikivoyage	9.02***	1.82	0.5169
Qwen2.5-32B	KOLIMO	6.27***	0.74	0.7731
	Wikivoyage	3.80*	1.50	0.3506
Mistral-Small-3.2-24B	KOLIMO	9.00***	0.93	0.7631
	Wikivoyage	5.96***	1.03	**0.5294**
Llama-3.2-3B	KOLIMO	7.10***	0.92	0.6658
	Wikivoyage	4.17***	1.05	0.3799

Table 4: Slopes from CLMs showing the **strength of alignment** (β-coefficient) between human and instruction-tuned LLM originality ratings in the two corpora. Stars denote significance (*$p < .05$, **$p < .01$, ***$p < .001$). Standard Error (SE) quantifies the uncertainty of the estimated coefficient. Spearman correlations are for direct comparison between LLMs.

MetaphOrig is thus among the first datasets combining approaches from psychological creativity research assessing originality holistically with approaches from linguistics and rhetoric that focus on specific, smaller textual units. Through the inclusion of LLM ratings, the datasets allow for systematic comparisons between human and LLM perceptions of originality.

A showcase study demonstrates how analyses can explicitly connect these perspectives in practice. It suggests that human ratings of originality can be traced back to metaphor density in a literary textual context and a similar pattern applies to LLM ratings. However, deliberateness of metaphor needs to be explored in future work, including conventionality and novelty of metaphor, as well as other stylistic factors and historical distance. Such analyses can draw on our small datasets and build upon the showcase study presented in this paper.

Limitations

The model hyperparameters for the prompting experiments were not individually optimized. Adjusting parameters such as the model temperature could potentially improve alignment with human ratings. Likewise, prompt design introduces variability. Different structuring of the prompts might have yielded different outcomes. Another limitation of this study is that it only considers the originality criterion of creativity. For a more comprehensive view on creativity, follow-up studies must take a closer look at the dimension of success as well. Moreover, we acknowledge that the provided contextual information (presenting the sentences of the KOLIMO dataset as everyday language) in the rating study may have shaped the ratings, for instance by making stylistic features that are conventional in literary texts appear more original. A follow-up study will therefore systematically investigate how different priming conditions and explicit genre labels affect originality ratings.

Acknowledgements

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A05.

Bibliographical References

Selcuk Acar. 2025. Creativity Assessment, Research, and Practice in the Age of Artificial Intelligence. *Creativity Research Journal*, 37(2):181–187. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2023.2271749.

Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013. Place: US Publisher: American Psychological Association.

Christian Burgers, Elly A. Konijn, Gerard J. Steen, and Marlies A.R. Iepsma. 2015. Making ads less complex, yet more creative and persuasive: the effects of conventional metaphors and irony in print advertising. *International Journal of Advertising*, 34(3):515–532.

Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. 2026. Evaluating text creativity across diverse domains: a dataset and large language model evaluator. In *The Fourteenth International Conference on Learning Representations*.

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, Toronto, Canada. Association for Computational Linguistics.

Rune H. B. Christensen. 2023. ordinal—Regression Models for Ordinal Data.

EAV Crompvoets. 2025. *Behind the scenes of pairwise comparison for educational measurement*. Ridderprint.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C. Neubauer. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1):35–40. Place: US Publisher: Educational Publishing Foundation.

Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-My Nguyen. 2024. Guidelines for the Annotation of Intentional Linguistic Metaphor. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Paul V. DiStefano, John D. Patterson, and Roger E. Beaty. 2024. Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2024.2326343.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Lettie Dorst. 2015. More or different metaphors in fiction? A quantitative cross-register comparison. *Language and Literature*, 24:3–22.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

B.A. Hennessey, Teresa M. Amabile, and J.S. Mueller. 2011. Consensual Assessment. In *Encyclopedia of Creativity*, pages 253–260. Elsevier.

J. Berenike Herrmann. 2015. High on metaphor, low on simile? An examination of metaphor type in sub-registers of academic prose. In J. Berenike Herrmann and Tony Berber Sardinha, editors, *Metaphor in Language, Cognition, and Communication*, volume 4, pages 163–190. John Benjamins Publishing Company, Amsterdam.

J. Berenike Herrmann and Gerhard Lauer. 2018. Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Osnabrücker Beiträge zur Sprachtheorie*, 92:127–156.

J. Berenike Herrmann, Karola Woll, and Aletta G. Dorst. 2019. Linguistic metaphor identification in German. *Metaphor Identification in Multiple Languages. MIPVU around the world*. ISBN: 9789027204721.

Rebecca M. M. Hicke and Ross Deans Kristensen-McLachlan. 2024. SCIENCE IS EXPLORATION: Computational Frontiers for Conceptual Metaphor Theory. In *CHR 2024: Computational Humanities Research Conference,*, pages 1105–1116, Aarhus, Denmark.

Jan Horstmann. 2019. KOLIMO: Korpus der literarischen Moderne. *forTEXT. Literatur digital erforschen*.

Jan Horstmann and E Akazawa. 2024. Ressourcenbeitrag: Korpus der literarischen Moderne (KOLIMO). *forTEXT, Korpusbildung*, 1(2). Medium: PDF,XML Publisher: Universitäts- und Landesbibliothek Darmstadt.

James C. Kaufman, John Baer, and Jason C. Cole. 2009. Expertise, Domains, and the Consensual Assessment Technique. *The Journal of Creative Behavior*, 43(4):223–233. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2162-6057.2009.tb01316.x.

Sungeun Kim and Dongsuk Oh. 2025. Evaluating Creativity: Can LLMs Be Good Evaluators in Creative Writing Tasks? *Applied Sciences*, 15(6):2971. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Christine A. Knoop and Stefan Blohm. 2025. Literarinesses—A Bag of Three-Sided Coins. *Literature*, 5(3):21.

Christine A. Knoop, Thomas Nehrlich, Sabrina Aristei, Oliver Lubrich, Kirsten Stark, Alexander Enge, Werner Sommer, and Rasha Abdel Rahman. 2024. The usual miracles: How narrative style affects the processing of counterintuitive concepts. *Psychology of Aesthetics, Creativity, and the Arts*.

Katrin Kohl, Marianna Bolognesi, and Ana Werkmann Horvat. 2020. The Creative Power of Metaphor. In Katrin Kohl, Rajinder Dudrah, Andrew Gosler, Suzanne Graham, Martin Maiden, and Wen-chin Ouyang, editors, *Creative Multilingualism*, 1 edition, pages 25–46. Open Book Publishers, Cambridge, UK.

Zoltán Kövecses. 2010. *Metaphor: a practical introduction*, 2. edition edition. Oxford University Press, Oxford.

Zoltán Kövecses. 2020. Conceptual metaphor theory. In Elena Semino and Zsófia Demjén, editors, *The Routledge handbook of metaphor and language*, first issued in paperback edition, Routledge handbooks in linguistics, pages 13–27. Routledge, Taylor & Francis Group, London New York.

George Lakoff and Mark Johnson. 2003. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Antonio Jr. Laverghetta, Tuhin Chakrabarty, Tom Hope, Jimmy Pronchick, Krupa Bhawsar, and Roger E. Beaty. 2025. How do Humans and Language Models Reason About Creativity? A Comparative Analysis. ArXiv:2502.03253 [cs].

Simone A. Luchini, Ibraheem Muhammad Moosa, John D. Patterson, Dan Johnson, Matthijs Baas, Baptiste Barbot, Iana Bashmakova, Mathias Benedek, Qunlin Chen, Giovanni E. Corazza, Boris Forthmann, Benjamin Goecke, Sameh Said-Metwaly, Maciej Karwowski, Yoed N. Kenett, Izabela Lebuda, Todd Lubart, Kirill G. Miroshnik, Felix-Kingsley Obialo, Marcela Ovando-Tellez, Ricardo Primi, Rogelio Puente-Díaz, Claire Stevenson, Emmanuelle Volle, Aleksandra Zielińska, Janet G. van Hell, Wenpeng Yin, and Roger E. Beaty. 2025. Automated assessment of creativity in multilingual narratives. *Psychology of Aesthetics, Creativity, and the Arts*. Place: US Publisher: Educational Publishing Foundation.

Rowan Hall Maudslay and Simone Teufel. 2022. Metaphorical Polysemy Detection: Conventional Metaphor meets Word Sense Disambiguation. ArXiv:2212.08395 [cs].

Mistral AI. 2025. Mistral-Small-3.2-24B-Instruct-2506.

Omar Momen, Emilie Sitter, Berenike Herrmann, and Sina Zarrieß. 2026. Surprisal and metaphor novelty: Moderate correlations and divergent scaling effects. *arXiv preprint arXiv:2601.02015*.

Seyedeh Hamideh Mozaffari. 2013. An Analytical Rubric for Assessing Creativity in Creative Writing. *Theory and Practice in Language Studies*, 3.

Andreas Nolda. 2024. Wikivoyage-Korpus: Korpusquellen der deutschen Sprachversion von Wikivoyage im TEI-Format.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Meihua Qian and Jonathan A Plucker. 2017. Creativity assessment. In *Creativity and innovation*, pages 223–234. Routledge.

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. 2024. Do LLMs Agree on the Creativity Evaluation of Alternative Uses? ArXiv:2411.15560 [cs].

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A Method for Identifying Potentially Deliberate Metaphor in Language Use. *Corpus Pragmatics*, 2(2):129–147.

Sebastian Reimann and Tatjana Scheffler. 2024. When is a Metaphor Actually Novel? Annotating Metaphor Novelty in the Context of Automatic Metaphor Detection. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2025. Using Large Language Models to Perform MIPVU-Inspired Automatic Metaphor Detection. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(1):92–96. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2012.650092.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in Spanish. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 45–56, Vienna, Austria. Association for Computational Linguistics.

Elena Semino and Gerard J. Steen. 2008. Metaphor in Literature. In Jr. Gibbs, Raymond W., editor, *The Cambridge Handbook of Metaphor*

and Thought, Cambridge Handbooks in Psychology, pages 232–246. Cambridge University Press, Cambridge.

Emilie Sitter, Omar Momen, Florian Steig, J. Berenike Herrmann, and Sina Zarrieß. 2025. Annotating Spatial Descriptions in Literary and Non-Literary Text. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 308–325, Vienna, Austria. Association for Computational Linguistics.

Gerard J. Steen. 2017. Deliberate Metaphor Theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14(1):1–24. Publisher: De Gruyter Mouton.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, and Tina Krennmayr. 2010a. Metaphor in usage. 21(4):765–796. Publisher: De Gruyter Mouton Section: Cognitive Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010b. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research (CELCR)*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Theresa J. Weinstein, Simon Majed Ceh, Christoph Meinel, and Mathias Benedek. 2022. What's Creative About Sentences? A Computational Approach to Assessing Creativity in a Sentence Generation Task. *Creativity Research Journal*, 34(4):419–430. Publisher: Routledge _eprint: https://doi.org/10.1080/10400419.2022.2124777.

Hadley Wickham, Thomas Lin Pedersen, and Dana Seidel. 2011. scales: Scale Functions for Visualization. Series Title: CRAN: Contributed Packages.

Wikimedia Foundation Inc. 2025. Wikivoyage – Freie Reiseinformationen rund um die Welt.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. ArXiv:2407.10671 [cs].

Fengying Ye, Shanshan Wang, Lidia S. Chao, and Derek F. Wong. 2025. Unveiling LLMs' Metaphorical Understanding: Exploring Conceptual Irrelevance, Context Leveraging and Syntactic Influence. ArXiv:2510.04120 [cs].

Claire M. Zedelius, Caitlin Mills, and Jonathan W. Schooler. 2019. Beyond subjective judgments: Predicting evaluations of creative writing from computational linguistic features. *Behavior Research Methods*, 51(2):879–894.

A. Rater Instructions

A.1. Original German Instruction

Wie originell und überraschend ist diese Beschreibung für Sie? Wird in der Textstelle ein Ort so beschrieben, wie Sie es in einem alltagssprachlichen Text nicht erwartet hätten? Gibt es eine frische Perspektive oder einen unerwarteten oder neuen Ausdruck?

Originelle Beschreibungen müssen nicht unbedingt extrem kunstvoll und ausgeschmückt sein. Entscheidend ist, ob Sie überrascht sind darüber, dass jemand einen Ort auf eine solche Art und Weise beschreibt.

Beispiele:

wenig originell: *Am frühen Morgen lag die Straße, die durch einige abgelegene Orte führte, im Nebel.*

sehr originell: *Die Straße verlief im Nebel des frühen Morgens und machte sich die Mühe, winzige Orte zu besuchen, die streng genommen nicht auf ihrem Weg lagen.*

A.2. English Translation

How original or surprising is this description for you? Does it describe a place in a way you would not have expected in everyday language? Is there a fresh perspective or an unexpected or new phrasing?

Original descriptions are not necessarily very elaborate and ornate. What is relevant is whether you are surprised about someone describing a place in such a way.

Examples:

not very original: *Early in the morning, the road that led through some remote places lay in the fog.*

very original: *The road ran away in the mist of the early morning, going to some trouble to visit tiny towns which were not, strictly speaking, on its way.*

B. LLM Prompts

B.1. Original German Prompt

System Prompt

> Du bist Annotator auf Prolific mit Deutsch als Erstsprache.
>
> Du nimmst an der Studie ‘Verstehen von Beschreibungen’ teil und bewertest Beschreibungen auf auf einer Skala von 1 bis 6.
>
> Die Beschreibungen stammen aus Kontexten des alltäglichen Lebens, wie Postkarten, Textnachrichten oder Wikipedia. Alle Beschreibungen haben die Gemeinsamkeit, dass sie räumliche Gegebenheiten beschreiben.

User Prompt

> Wie originell und überraschend ist diese Beschreibung für Sie? Wird in der Textstelle ein Ort so beschrieben, wie Sie es in in einem alltagssprachlichen Text nicht erwartet hätten? Gibt es eine frische Perspektive oder einen unerwarteten oder neuen Ausdruck?
>
> Originelle Beschreibungen müssen nicht unbedingt extrem kunstvoll und ausgeschmückt sein. Entscheidend ist, ob Sie überrascht sind darüber, dass jemand einen Ort auf eine solche Art und Weise beschreibt.
>
> **Beispiele:**
>
> wenig originell: *Am frühen Morgen lag die Straße, die durch einige abgelegene Orte führte, im Nebel.*
>
> sehr originell: *Die Straße verlief im Nebel des frühen Morgens und machte sich die Mühe, winzige Orte zu besuchen, die streng genommen nicht auf ihrem Weg lagen.*
>
> **Beschreibung zur Bewertung:** {sentence}
>
> Bewerte den Satz mit einer Zahl zwischen 1 und 6 und begründe deine Entscheidung. Antworte ausschließlich im JSON-Format mit folgenden Feldern: ‘Bewertung’ und ‘Begründung’.

B.2. English Translation

System Prompt

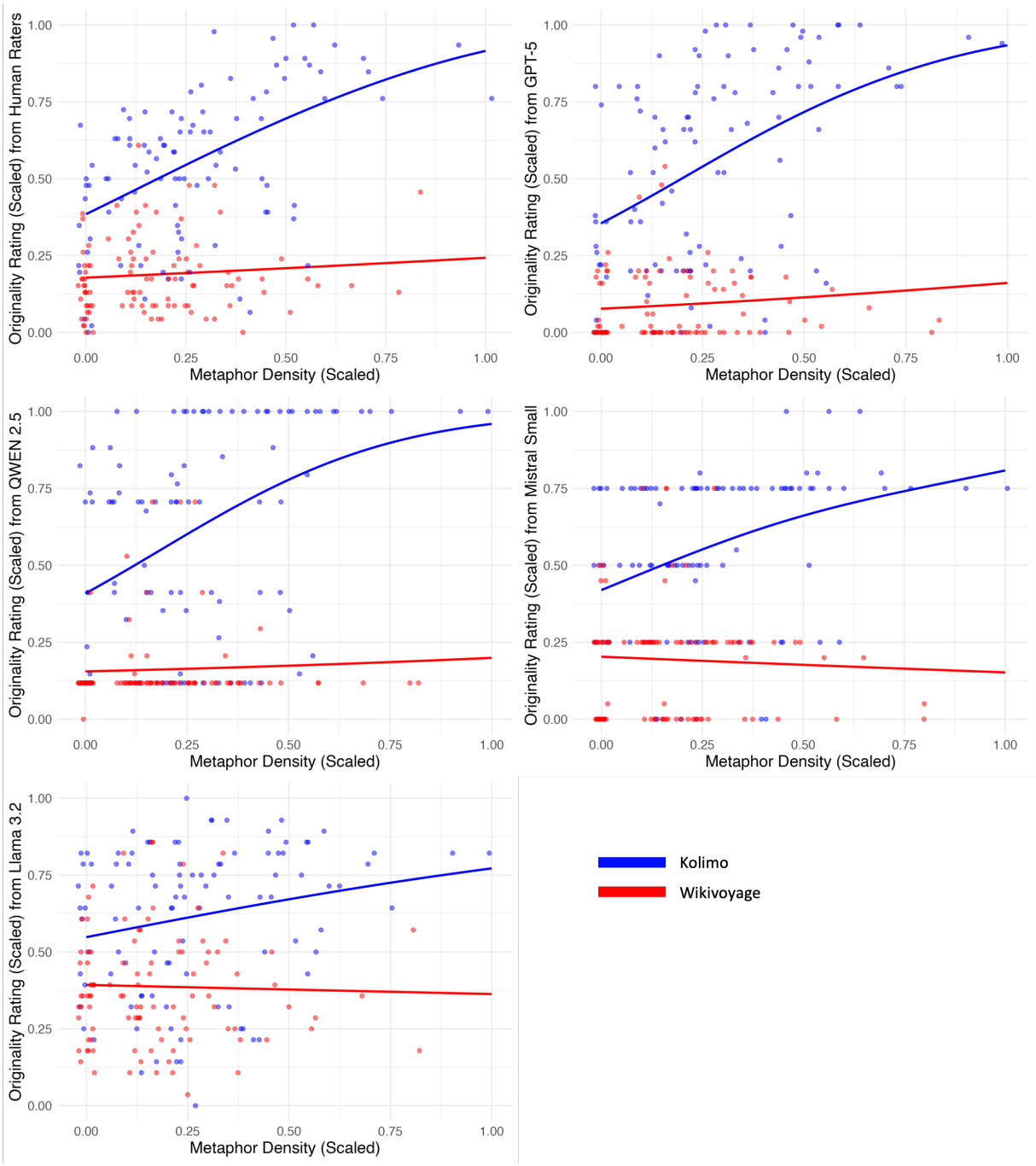
> You are an German native speaker annotator on Prolific.
>
> You are participating in the study ‘Understanding Descriptions’ and evaluating descriptions on a scale from 1 to 6.
>
> The descriptions come from everyday contexts, such as postcards, text messages, or Wikipedia. All descriptions have in common that they describe spatial surroundings.

User Prompt

> How original and surprising is this description to you? Does the passage describe a place in a way that you would not expect in everyday language? Is there a fresh perspective or an unexpected or new expression?
>
> Original descriptions do not necessarily have to be extremely artistic and embellished. What is relevant is whether you are surprised that someone would describe a place in such a way.
>
> **Examples:**
>
> not very original: *Early in the morning, the road that led through some remote places lay in the fog.*
>
> very original: *The road ran away in the mist of the early morning, going to some trouble to visit tiny towns which were not, strictly speaking, on its way.*
>
> **Description for evaluation:** {sentence}
>
> Rate the sentence with a number between 1 and 6 and explain your decision. Respond exclusively in JSON format with the following fields: ‘rating’ and ‘reason’.

C. Averaged ratings per sentence and ordinal regression lines

Increase/decrease of lines for the Wikivoyage dataset is not significant.


Originality Rating (Scaled) from Human Raters
Originality Rating (Scaled) from GPT-5
Originality Rating (Scaled) from QWEN 2.5
Originality Rating (Scaled) from Mistral Small
Originality Rating (Scaled) from Llama 3.2
Metaphor Density (Scaled)
0.00
0.25
0.50
0.75
1.00
Kolimo
Wikivoyage

Author Index