



LREC 2026

NLP4Ecology @ LREC 2026

Workshop Proceedings

Editors

**Francesca Grasso, Valerio Basile, Cristina Bosco,
Muhammad Okky Ibrohim, Maria Skeppstedt, and
Manfred Stede**

12 May 2026

NLP4Ecology 2026

The 2nd Workshop on Ecology, Environment, and Natural Language Processing

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-53-1

Preface

We are pleased to welcome you to NLP4Ecology 2026, the 2nd edition of the International Workshop on Ecology, Environment, and Natural Language Processing. Following a successful first edition, which confirmed a rapidly growing community interested in the intersection of NLP and ecology, this year's workshop takes place on May 12th, 2026, co-located with LREC 2026 in Palma de Mallorca, Spain. The larger venue and its broader audience provide an important opportunity to further expand the visibility and impact of this initiative.

NLP4Ecology is designed to be a venue for publication and exchange between the Natural Language Processing (NLP) community and researchers and practitioners from a wide range of disciplines. It aims to explore how computational linguistics and NLP tools, methods, and applications can contribute to addressing urgent environmental issues - not only climate change, which has received the most research attention so far due to its visibility and quantifiability, but also broader ecological crises affecting humans, non-human animals, and ecosystems worldwide. Tackling these issues requires interdisciplinary collaboration, and the NLP community has an important role to play in this effort.

The workshop continues to stimulate an interdisciplinary space at the intersection of computational linguistics and environmental and ecological research. The contributions in this edition are primarily rooted in NLP and applied computational methods, while also attracting (eco-)linguistic perspectives on environmental discourse, including the analysis of narratives, constructions, and framing. In this context, the workshop aims to support the development of new research directions and collaborations, and to encourage critical reflection on the role of language technologies in actively shaping environmental knowledge and discourse.

This year's program reflects both continuity and growth. We received a total of 22 submissions, all of which underwent a double-blind review process. In the end, 14 papers were accepted, corresponding to an overall acceptance rate of 63%. We are deeply grateful to the members of our Program Committee for their careful and constructive evaluations, which were essential in shaping a diverse and high-quality program.

The workshop features a keynote lecture by **Leonie Bossert**, whose work at the intersection of environmental ethics and philosophy of technology offers an enlightening perspective on questions of responsibility, technology, and human-nature relations.

The accepted papers reflect a range of approaches to environmental language and data. Several contributions focus on sustainability communication and ESG reporting, addressing challenges such as greenwashing detection, readability, and alignment with reporting standards. Others examine environmental discourse and narratives, including climate disinformation, media coverage of extreme events, and the analysis of historical and multilingual corpora. A number of papers investigate the role of large language models in ecological contexts, with particular attention to evaluation, interpretability, and explanatory capabilities, as well as their alignment with ecological values. Additional contributions address applied NLP tasks, including corpus construction in lower-resource settings, biodiversity data processing, and retrieval-based approaches to environmental information. The contributions draw on a variety of data sources, including social media content, news corpora, sustainability and ESG reports, biodiversity databases, and historical travel narratives. This diversity of data highlights the wide range of contexts in which environmental language is produced and interpreted, and supports a more comprehensive understanding of how ecological issues are communicated across domains.

Regarding linguistic and geographical diversity, the accepted papers reflect a broad range of languages and research contexts, including English, German, Russian, and Vietnamese, as

well as multilingual evaluation settings. The contributing authors are affiliated with institutions across multiple countries, including Italy, Germany, the Netherlands, the United Kingdom, Spain, Greece, Cyprus, and the Philippines, among others.

Organizing a workshop of this kind requires the support and collaboration of many people. We would like to sincerely thank our Program Committee members again for their time and expertise, as well as all authors for their submissions and contributions. Finally, we thank all participants for being part of this second edition and for helping to build a growing research community at the intersection of NLP and ecology.

NLP4Ecology 2026 Organizing Committee

Francesca Grasso
Valerio Basile
Cristina Bosco
Muhammad Okky Ibrohim
Maria Skeppstedt
Manfred Stede

May 2026

Organizing Committee

Workshop Chairs

- Francesca Grasso, University of Turin, Italy
- Valerio Basile, University of Turin, Italy
- Cristina Bosco, University of Turin, Italy
- Muhammad Okky Ibrohim, Universitas Indonesia, Indonesia, and University of Tartu, Estonia
- Maria Skeppstedt, Uppsala University, Sweden
- Manfred Stede, University of Potsdam, Germany

Program Committee

- Luca Brigada Villa, University of Pavia, Italy
- Tyler A. Chang, University of California San Diego; Google DeepMind, USA
- Giacomo De Luca, University of Rome "Tor Vergata", Italy
- Steffen Frenzel, University of Potsdam, Germany
- Sara Gemelli, University of Bergamo, Italy
- Anna Glazkova, University of Tyumen, Russia
- Amanda Starling Gould, Duke University, USA
- Frederik Hudi, Nara Institute of Science and Technology, Japan
- Veronika Laippala, University of Turku, Finland
- Stefano Locci, University of Turin, Italy
- Stella Markantonatou, Athena Research Center, Greece
- Luca Moroni, Sapienza University of Rome, Italy
- Robert Östling, Stockholm University, Sweden
- Lucia Passaro, University of Pisa, Italy
- Melanie Siegel, Darmstadt University of Applied Sciences, Germany
- Maximos Skandalis, CNRS & University of Montpellier, France
- Ivan Spada, Fondazione Bruno Kessler (FBK), Italy
- Masashi Takeshita, Nagoya University, Japan
- Fabio Massimo Zanzotto, University of Rome "Tor Vergata", Italy

Invited Speaker

- Leonie N. Bossert, University of Vienna, Austria

Table of Contents

<i>Retrieving Climate Change Disinformation by Narrative</i> Max Upravitelev, Veronika Solopova, Charlott Jakob, Premtim Sahitaj, Sebastian Möller and Vera Schmitt	1
<i>Unsupervised GRI-TCFD Alignment with LLM-Assisted Validation for Climate Disclosure and Greenwashing Risk Analysis</i> Seyed Alireza Mousavian Anaraki, Danilo Croce, Roberta Costa, Luigi Tiburzi, Armando Calabrese and Roberto Basili	15
<i>Towards Empowering Consumers through Sentence-level Readability Scoring in German ESG Reports</i> Benjamin Josef Schüßler and Jakob Prange	26
<i>Disambiguating Geographic Names in Biodiversity Occurrence Data: A Retrieval-Augmented Generation Approach</i> Yanni Jose C. Ella, Monica Ashley R. Laviste, John Michael L. Lastimoso, Wilfred John E. Santiañez, Riza Batista-Navarro and Roselyn Santos Gabud	42
<i>Sentiment and Stance in EFL Responses to AI-Generated Environmental Content</i> Andry Sophocleous	53
<i>What Stories Do Language Models Tell About Nature? A Multi Layer Evaluation Framework for Ecological Alignment</i> Jorge Vallego, Eleanor Tiernan, Mah Rukh, Mariana Roccia and Sabina Fiebig Lord ...	60
<i>Ecological Discourse Modeling in a Low-Resource Setting: A Longitudinal Vietnamese Climate Corpus with Comparative Topic Modeling</i> Huyen Phuong Nguyen	69
<i>Greench-v1: distilling SLMs on Greenwashing Detection</i> Federico Raspanti, Alessandro Pietro Bardelli Bardelli, Simona Scala, İrem Demirtaş, Mar- ilena Di Bari and Michele Filannino	79
<i>Analyzing Environmental Discourse through Construction-Based Pattern Extraction</i> Elisa Chierchiello, Eliana Di Palma, Ludovica Pannitto and Cristina Bosco	87
<i>Mapping the Historical Ecology of the Cyclades: A Diachronic Natural Language Processing Analysis of Travel Narratives (1700–1920)</i> Aikaterini Christopoulou, Vassilis Detsis and Basilis Gatos	105
<i>Retrieving Floods without Floodlights: Topic Models as Binary Classifiers for Extreme Climate Events in German News</i> Brielen Madureira, Mariana Madruga de Brito and Andreas Niekler	115
<i>Why Is This Green? LLM-Based Explanations of Implicit Green Practices in Social Media</i> Anna Glazkova, Olga Zakharova and Daria Lebedeva	135
<i>Introducing a Green Leaderboard for Sustainable Risk Prediction in Streaming NLP Shared Tasks.</i> Alba María Mármol-Romero, Adrián Moreno Muñoz and Arturo Montejo-Raez	144

Not Everything Is Greenwashing: Limitations of Automatic Analysis of Sustainability Reports, and a Proposal
Maria Pilar Uribe Silva, Rik van Noord and Malvina Nissim 154

Workshop Program

Tuesday, May 12, 2026

- 14:10–14:55** **Invited Talk: Leonie N. Bossert**
- 14:55–15:55** **Session 1**
- 14:55–15:10 *Retrieving Climate Change Disinformation by Narrative*
Max Upravitelev, Veronika Solopova, Charlott Jakob, Premtim Sahitaj,
Sebastian Möller and Vera Schmitt
- 15:10–15:25 *Unsupervised GRI-TCFD Alignment with LLM-Assisted Validation for
Climate Disclosure and Greenwashing Risk Analysis*
Seyed Alireza Mousavian Anaraki, Danilo Croce, Roberta Costa, Luigi
Tiburzi, Armando Calabrese and Roberto Basili
- 15:25–15:40 *Towards Empowering Consumers through Sentence-level Readability
Scoring in German ESG Reports*
Benjamin Josef Schübler and Jakob Prange
- 15:40–15:55 *Disambiguating Geographic Names in Biodiversity Occurrence Data: A
Retrieval-Augmented Generation Approach*
Yanni Jose C. Ella, Monica Ashley R. Laviste, John Michael L. Lasti-
moso, Wilfred John E. Santiañez, Riza Batista-Navarro and Roselyn
Santos Gabud
- 16:00–16:50** **Poster Session**
- 16:00–16:50 *Sentiment and Stance in EFL Responses to AI-Generated Environmen-
tal Content*
Andry Sophocleous
- 16:00–16:50 *What Stories Do Language Models Tell About Nature? A Multi Layer
Evaluation Framework for Ecological Alignment*
Jorge Vallego, Eleanor Tiernan, Mah Rukh, Mariana Roccia and Sabina
Fiebig Lord
- 16:00–16:50 *Ecological Discourse Modeling in a Low-Resource Setting: A Longitu-
dinal Vietnamese Climate Corpus with Comparative Topic Modeling*
Huyen Phuong Nguyen
- 16:00–16:50 *Greench-v1: distilling SLMs on Greenwashing Detection*
Federico Raspanti, Alessandro Pietro Bardelli Bardelli, Simona Scala,
İrem Demirtaş, Marilena Di Bari and Michele Filannino

- 16:00–16:50 *Analyzing Environmental Discourse through Construction-Based Pattern Extraction*
Elisa Chierchiello, Eliana Di Palma, Ludovica Pannitto and Cristina Bosco
- 16:00–16:50 *Mapping the Historical Ecology of the Cyclades: A Diachronic Natural Language Processing Analysis of Travel Narratives (1700–1920)*
Aikaterini Christopoulou, Vassilis Detsis and Basilis Gatos
- 17:00–17:55 **Session 2****
- 17:00–17:15 *Retrieving Floods without Floodlights: Topic Models as Binary Classifiers for Extreme Climate Events in German News*
Brielen Madureira, Mariana Madruga de Brito and Andreas Niekler
- 17:15–17:25 *Why Is This Green? LLM-Based Explanations of Implicit Green Practices in Social Media*
Anna Glazkova, Olga Zakharova and Daria Lebedeva
- 17:25–17:40 *Introducing a Green Leaderboard for Sustainable Risk Prediction in Streaming NLP Shared Tasks.*
Alba María Mármol-Romero, Adrián Moreno Muñoz and Arturo Montejor-Raez
- 17:40–17:55 *Not Everything Is Greenwashing: Limitations of Automatic Analysis of Sustainability Reports, and a Proposal*
Maria Pilar Uribe Silva, Rik van Noord and Malvina Nissim

Retrieving Climate Change Disinformation by Narrative

Max Upravitelev^{1,2}, Veronika Solopova^{1,2}, Charlott Jakob^{1,2},
Premtim Sahitaj^{1,2}, Sebastian Möller^{1,2} and Vera Schmitt^{1,2,3,4}

¹Technische Universität Berlin, ²German Research Center for Artificial Intelligence (DFKI)

³BIFOLD – Berlin Institute for the Foundations of Learning and Data

⁴Centre for European Research in Trusted AI (CERTAIN)

max.upravitelev@tu-berlin.de

Abstract

Detecting climate disinformation narratives typically relies on fixed taxonomies, which do not accommodate emerging narratives. Thus, we re-frame narrative detection as a retrieval task: given a narrative’s core message as a query, rank texts from a corpus by alignment with that narrative. This formulation requires no predefined label set and can accommodate emerging narratives. We repurpose three climate disinformation datasets (CARDS, Climate Obstruction, climate change subset of PolyNarrative) for retrieval evaluation and propose SpecFi, a framework that generates hypothetical documents to bridge the gap between abstract narrative descriptions and their concrete textual instantiations. SpecFi uses community summaries from graph-based community detection as few-shot examples for generation, achieving a MAP of 0.494 on CARDS without access to narrative labels. We further introduce narrative variance, an embedding-based difficulty metric, and show via partial correlation analysis that standard retrieval degrades on high-variance narratives (BM25 loses 63.4% of MAP), while SpecFi-CS remains robust (32.7% loss). Our analysis also reveals that unsupervised community summaries converge on descriptions close to expert-crafted taxonomies, suggesting that graph-based methods can surface narrative structure from unlabeled text.

Keywords: information retrieval, disinformation detection, climate change denial

1. Introduction

Recent datasets on climate change disinformation (Coan et al., 2021), (Nikolaidis et al., 2025), (Rowlands et al., 2024) organize individual claims under narrative taxonomies defined by core messages. These taxonomies group together texts, sometimes with little lexical overlap: the claim “Carbon dioxide is vital to all life on Earth because no vegetation can exist without it” and a lengthy scientific rebuttal arguing that “the historical increase in the atmosphere’s CO₂ concentration has been good for the Amazon’s trees” share near-zero Jaccard similarity, yet both serve the same narrative: that CO₂ is essentially plant food (a narrative from the CARDS taxonomy (Coan et al., 2021)).

Framing narrative identification as classification enables the detection of known narratives but limits adaptability: classification assumes a fixed label set, whereas disinformation narratives evolve. Re-framing the task as retrieval, where a narrative’s core message serves as a query to rank candidate texts, enables a more flexible monitoring strategy that can target emerging, previously unseen narratives. In practice, this means that when fact-checkers or journalists observe a potentially emerging narrative, they could formulate its core message as a query and search a corpus to assess how prevalent it already is without requiring a predefined label set or retraining a classifier. However, this flexibility comes at a cost: narrative retrieval poses its own challenges. Unlike standard semantic search, which matches surface-level meaning,

narrative retrieval must identify texts by their underlying core message, which may never be stated explicitly. Prior work has shown that dense retrievers fail to respect implicit logical constraints in queries (Shen et al., 2025) and that text embedding models struggle with structural and relational understanding between concepts (Zhang et al., 2023). Narrative understanding specifically remains a known limitation of current language models (Zhu et al., 2023). These failures cascade in narrative retrieval, where queries express abstract core messages (e.g., “CO₂ is plant food”) that texts may support through implicit logical entailment or varied syntactic framings without stating the theme directly: the difficulty is the gap between narrative descriptions, which are abstract, and their textual instantiations, which are concrete and stylistically diverse.

In this paper, we explore narrative retrieval in the domain of climate disinformation. Our primary contributions are analytical rather than architectural: the individual components of our pipeline, including dense retrieval, dynamic few-shot sampling, HyDE-style generation (Gao et al., 2023), and graph-based community detection via the framework NodeRAG (Xu et al., 2025), are drawn from existing work. Their combination serves as the experimental setup for three contributions:

1. **Retrieval-based evaluation of narrative datasets.** We repurpose three climate disinformation narrative datasets (CARDS, Climate Obstruction, a climate change-related subset of PolyNarrative) for retrieval evaluation, using narrative labels as queries and associated texts

as relevance judgments.

- 2. SpecFi: Speculative Fiction generation for narrative retrieval.** We propose a framework that bridges the gap between abstract narrative descriptions and concrete textual instantiations by generating hypothetical documents following the HyDE strategy (Gao et al., 2023). SpecFi¹ operates in two variants: SpecFi-DR retrieves the nearest text from the reference corpus via dense retrieval as a few-shot example. SpecFi-CS retrieves high-level community summaries via graph-based search over a heterogeneous knowledge graph (Xu et al., 2025). Our evaluation shows that the community summaries improve performance beyond what actual samples from the training set achieve. Our analysis further reveals that these summaries can converge on descriptions close to expert-crafted narrative taxonomies like CARDS (Coan et al., 2021), suggesting that graph-based methods can extract narrative structure from unlabeled text; a property with application for monitoring emerging narratives.
- 3. Narrative variance as a predictor of retrieval difficulty.** We propose narrative variance (V_i), an embedding-based metric quantifying the internal spread of texts within a narrative group, and show via partial correlation analysis (controlling for sample size) that it correlates with retrieval difficulty for standard sparse and dense baselines. SpecFi-CS shows the smallest sensitivity to this effect, maintaining stable performance across high-variance narratives.

2. Preliminaries and Related Work

2.1. Disinformation Narrative Classification and Retrieval

Several recently released works organize disinformation texts under narrative taxonomies on different topics (Kotseva et al., 2023; Sosnowski et al., 2024; Haouari et al., 2025; Heinrich et al., 2024). Our focus in this paper is specific to climate change denial narrative datasets (CARDS (Coan et al., 2021)), climate obstruction in social media advertising (CO, (Rowlands et al., 2024)), and climate disinformation in news (PolyNarrative, (Nikolaidis et al., 2025), which consists of two topic splits: Climate Change and War in Ukraine).

Within related domains, the term “narrative retrieval” is used mainly to describe claim retrieval in practice, focusing on individual claims, not overarching elements like core messages, such as in Singh et al. (2024); Singh (2024). Akter and Santu (2024) identified the need for metrics that capture

narrative similarity beyond surface representations, and Hatzel and Biemann (2024) demonstrated the difficulty of narrative retrieval by showing that untailored dense retrieval substantially underperforms on the task of retrieving texts by their summaries.

Hypothetical Document Embeddings The retrieval strategy of generating hypothetical documents to bridge the gap between query and document representations was introduced as HyDE by Gao et al. (2023). Given a query, HyDE generates n hypothetical documents, embeds them, and retrieves based on the aggregated embeddings. This effectively expands queries with vocabulary and semantic meaning which is representative for relevant documents, which is a valuable property for narrative retrieval, where narrative descriptions are abstract while their instantiations are concrete. This generation step can be understood as a computational analogue of what Roine (2020) calls the instrumental mode of speculation: generating possible consequences from a given premise. We adopt this framing in our system name (SpecFi, Speculative Fiction).

Graph-Based Reasoning in Retrieval Narratives can often be understood as sets of narrative features and their interrelated structures (Piper et al., 2021; Hellman, 2024). Since embedding-based similarity can fail at capturing complex relational structures (as discussed in the introduction) graph-based representations offer an alternative: they can explicitly model entities, relationships, and thematic co-occurrence patterns. Within current retrieval research, several recent graph-based RAG frameworks construct knowledge graphs from unstructured corpora and apply community detection to identify thematic clusters. GraphRAG (Edge et al., 2025) introduced this paradigm: an LLM extracts entities and relationships, the Leiden algorithm (Traag et al., 2019) partitions the resulting graph into hierarchically nested communities, and a second LLM pass generates bottom-up summaries for each community. These summaries serve as coarse semantic layers for query-focused summarization at retrieval time. We build on NodeRAG (Xu et al., 2025), which refines this approach by operating over a heterogeneous graph with a search pipeline that propagates relevance through the graph structure; details are given in section 3.1.

3. Methodology

3.1. Retrieval Pipeline

The retrieval pipeline, illustrated in Figure 1, operates in five steps. We model a narrative monitoring scenario in which an analyst queries a text corpus

¹Reference code is available at: <https://github.com/XplainNLP/SpecFi-Narrative-Retrieval>

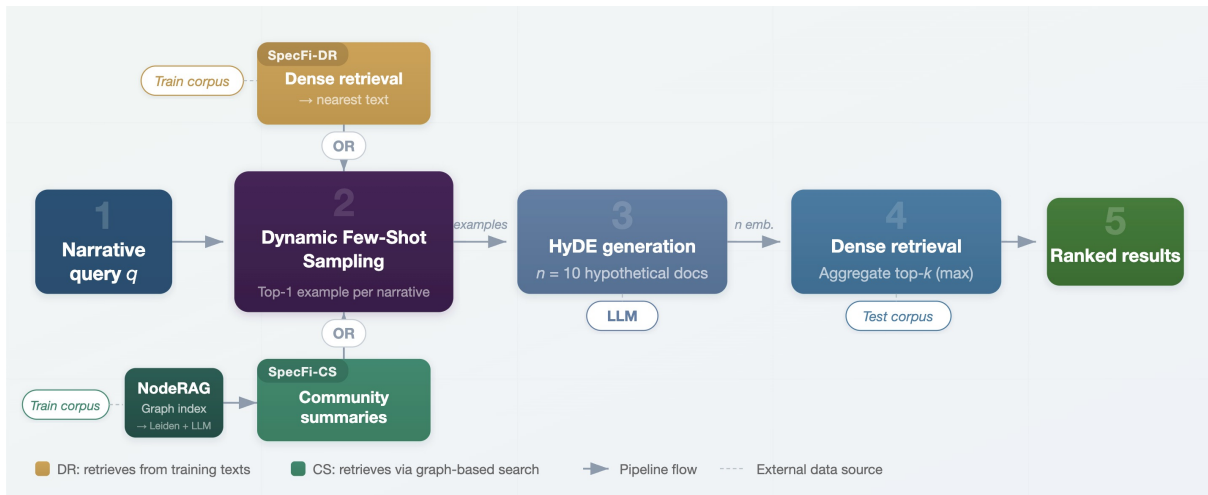


Figure 1: Overview: The SpecFi Retrieval Pipeline

by core message to identify texts aligned with a given narrative. For each dataset, we use the training split as a reference corpus and the test split as the evaluation set; narrative descriptions derived from each dataset’s taxonomy serve as query proxies (collected in our code repository). The reference corpus is used without access to narrative labels, simulating a realistic setting in which previously collected texts (including non-disinformation content) are available but lack narrative annotations.

Algorithm 1 SpecFi Narrative Retrieval Workflow

Require: Narrative taxonomy with labels used as queries $\{q_1, \dots, q_K\}$, reference corpus t , NodeRAG graph index \mathcal{G}

Ensure: Ranked list of candidate texts per narrative

- 1: Select target narrative q_k
- 2: For each q_k , retrieve one example via:
 - (a) SpecFi-DR: nearest text from t by cosine similarity, or
 - (b) SpecFi-CS: top-ranked high-level element from \mathcal{G} via NodeRAG’s graph-based search

and concatenate all K narrative-example pairs as few-shot examples

- 3: For target q_k , generate $n=10$ hypothetical documents
- 4: Embed hypotheticals; retrieve top- k from test set via aggregated dense retrieval
- 5: Return ranked results

NodeRAG (Xu et al., 2025) constructs a heterogeneous graph from the input corpus comprising seven node types, including entities, relationships, semantic units, and text chunks. During the graph augmentation stage, the Leiden community detec-

tion algorithm (Traag et al., 2019) is applied to segment the graph into communities of closely related nodes. For each detected community, an LLM analyzes the aggregated content of its member nodes and generates high-level element nodes which are represented by community summaries. These high-level elements are reintegrated into the graph, providing a summarization layer that captures patterns beyond what is present in any individual text. At query time, NodeRAG’s search combines embedding similarity and entity matching to identify seed nodes, then propagates relevance scores through the heterogeneous graph via Personalized PageRank. This means that a high-level element can be surfaced not only through direct similarity to the query but also through structural connectivity to other relevant nodes. In our SpecFi-CS pipeline, we query this search pipeline with each narrative description and extract the top-ranked high-level element from the retrieval results, using it as a few-shot example for hypothetical document generation. This exploits the summaries’ abstracted nature to produce hypotheticals that span the interpretive range of a narrative rather than anchoring on a single text. For each narrative, we generate $n=10$ hypothetical documents, selected based on preliminary experiments.

Illustrative Example Consider the CARDS narrative “Climate impacts / global warming is beneficial / not bad. CO₂ is beneficial / not a pollutant. CO₂ is plant food” (narrative id: 3_3).

SpecFi-DR retrieves the nearest text from the reference corpus as a few-shot example:

“Idso pointed out that there is a huge body of literature on the biological impacts of rising temperatures and atmospheric CO2 levels that the International Panel on Climate Change

(IPCC) ignores. He emphatically stated that atmospheric CO2 is not a pollutant. In fact, increased levels of CO2 reduce the negative effects of a number of plant stresses [...] and protects against herbivores.”

SpecFi-CS instead retrieves community summary:

“Some argue that the effects of CO2 increases and slight global warming may be harmless or even beneficial, challenging alarmist narratives about climate change.”

Notably, the community summaries are generated without access to narrative labels; we discuss their convergence with the expert-crafted taxonomy in section 6.

3.2. Datasets

CARDS The Computer-Assisted Recognition of (Climate Change) Denial and Skepticism dataset (Coan et al., 2021) contains climate change denial claims organized under a two-level taxonomy of 5 main narratives and 27 subnarratives, of which 17 are attested in the data. Each text is a short claim (mean 65 words) mapped to one narrative. With 2,904 texts in the test set and 21-225 texts per narrative, CARDS provides the densest evaluation setting and is the primary dataset for our statistical analysis.

Climate Obstruction (CO) The Climate Obstruction dataset (Rowlands et al., 2024) contains social media advertisements from oil and gas companies, classified under 7 obstruction narratives such as corporate community engagement and clean energy leadership. Here, the texts are shorter (mean 28 words), may carry multiple labels and are designed to reshape public perception of the fossil fuel industry.

PolyNarrative Climate Change Subset (PN-CC) The PolyNarrative dataset (Nikolaidis et al., 2025) contains news articles annotated with fine-grained narrative labels across multiple topics. For better comparability, we use the English language climate change related subset. Texts are substantially longer (mean 496 words) and frequently carry multiple narrative labels. With only 56 climate-related texts in the development set (used as test set; labels were not released for the actual test split), PN-CC serves as a complementary low-resource evaluation but does not support reliable statistical analysis.

The three datasets differ across several dimensions relevant to narrative retrieval evaluation, allowing us to test whether SpecFi generalizes across the heterogeneous landscape of climate disinformation. Table 1 summarizes quantitative statistics of the datasets.

Narrative descriptions used as queries are constructed from each dataset’s taxonomy by concatenating hierarchical labels (e.g., for CARDS: “Global warming is not happening. Ice/permafrost/snow cover isn’t melting”).

3.3. Metrics

Retrieval Performance We report Mean Average Precision (MAP), which summarizes precision across all recall levels; normalized Discounted Cumulative Gain at cutoffs 10 and 100 (nDCG@10, nDCG@100), which measures ranking quality with position-based discounting; and average R-Precision, the precision at the rank equal to the number of relevant documents. All are standard information retrieval metrics (Manning et al., 2008). Each evaluation is performed over K narratives per dataset, yielding K per-narrative scores that we aggregate by macro-averaging.

Embedding-Based Narrative Metrics Let $\mathcal{N} = \{n_1, \dots, n_K\}$ be a set of narratives. Each narrative n_i has an associated set of text embeddings $\mathcal{T}_i = \{\mathbf{t}_{i1}, \dots, \mathbf{t}_{im_i}\} \subset \mathbb{R}^d$ with centroid $\mathbf{c}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{t}_{ij}$. We define cosine distance as $d_{\cos}(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$.

Narrative Distinctness, as proposed in Irani et al. (2025), measures how separable a narrative is from the others via inter-centroid distances $d_{ij} = d_{\cos}(\mathbf{c}_i, \mathbf{c}_j)$. The geometric mean balances global separation (mean distance) with local distinctiveness (minimum distance):

$$D_i = \sqrt{\bar{d}_i \cdot d_i^{\min}}, \quad (1)$$

$$\bar{d}_i = \frac{1}{K-1} \sum_{j \neq i} d_{ij}, \quad d_i^{\min} = \min_{j \neq i} d_{ij}.$$

Narrative Variance measures the overall spread of texts around the centroid:

$$V_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \|\mathbf{t}_{ij} - \mathbf{c}_i\|_2^2. \quad (2)$$

The two metrics operationalize different aspects of the notion of measuring a narrative’s interpretation space: D_i captures how separable this space is from neighboring narratives and V_i captures the overall spread of instantiations around the narrative’s center. We treat them as competing hypotheses about what drives retrieval difficulty: is it proximity to other narratives (D_i) or overall internal diversity (V_i)? We test this in section 5.

3.4. Model Choice

For hypothetical document generation, we use gpt-4o (OpenAI, 2024) and gemma-3-27b-it (Team

	narratives <i>n</i>	mean texts per <i>n</i>	std texts per <i>n</i>	mean words per <i>n</i>	std words per <i>n</i>	mean words per text	std words per text	total texts	disinfo % of all texts
CARDS	17	67.65	57.49	7.61	3.96	65.35	57.60	2904	39.6
CO	7	38.29	27.20	20.50	3.56	28.27	11.95	255	73.3
PN	51	1.98	1.39	9.06	2.82	601.83	293.73	41	73.2
PN-UKR	27	2	1.47	9.32	3.30	740.62	382.17	13	100
PN-CC	23	2	1.32	8.77	2.12	495.71	120.74	17	100
PN-Neutral	0	-	-	-	-	459.78	166.97	11	0

Table 1: Quantitative statistics of the used datasets.

Setup	Models	MAP	NDCG @10	NDCG @100	Average R-Precision
zero shot	4o, OI-E	0.321	0.509	0.487	0.370
	G, Q4B	0.313	0.469	0.456	0.371
	G-a, Q4B	0.295	0.428	0.436	0.308
static*	4o, OI-E	0.488	0.713	0.649	0.487
	G, Q4B	0.435	0.635	0.616	0.435
	G-a, Q4B	0.464	0.679	0.637	0.468
SpecFi -DR	4o, OI-E	0.421	0.682	0.600	0.440
	G, Q4B	0.424	0.630	0.581	0.432
	G-a, Q4B	0.457	0.693	0.619	0.453
SpecFi -CS	4o, OI-E	0.426	0.660	0.597	0.456
	G, Q4B	0.468	0.709	0.631	0.492
	G-a, Q4B	0.494	0.726	0.657	0.487

Table 2: Results on the CARDS dataset. static* is included for reference only due to its reliance on labels. All metrics are averaged over 10 runs. We report a standard deviation of <0.01 for all performance metrics. The model abbreviations are: 4o=gpt-4o, OI-E= text-embedding-3-large, G=gemma-3-27b-it, G-a=gemma-3-27b-it ablated, Q4B=Qwen3-Embedding-4b.

et al., 2025) (including an uncensored or “ablated” variant with safety alignment removed in post-training to mitigate possible refusals when generating disinformation texts, denoted -a). The models are run as Q8_0 GGUF quantizations. For dense retrieval embeddings, we use Qwen3-Embedding-4B (Yang et al., 2025) due to its strong performance on MTEB² and support for instruction prompts. For the embedding-based narrative metrics (D_i, V_i), we use GTE-LARGE (Li et al., 2023) ($d = 1024$) to separate the analysis from the retrieval pipeline. NodeRAG graph construction follows the framework’s default configuration with OpenAI models for structured output generation.

4. Evaluation

Retrieval Performance We first evaluate our system on performance metrics to further analyze possible correlations with the narrative metrics introduced above. Table 2 documents our results,

²<https://huggingface.co/spaces/mteb/leaderboard>

where averages of metrics over 10 runs are presented due to randomized factors within HyDE. On CARDS, SpecFi-CS with the ablated model achieves the highest MAP (0.494) among all label-free setups, outperforming both the dense baseline (0.299) and SpecFi-DR (0.457). On CO, SpecFi-DR outperforms SpecFi-CS (0.519 vs. 0.491), suggesting that the relative advantage of community summaries over retrieved texts depends on dataset characteristics. For comparison, we also include the setups labeled with “static” where few-shot examples were statically retrieved by assessing the labels.

Setup/Model	MAP	NDCG @10	Avg. R-Prec.
BM25	0.326	0.472	0.298
Qwen3-E-4B	0.499	0.607	0.491
SpecFi-DR	0.519	0.644	0.496
SpecFi-DR-a	0.482	0.604	0.494
SpecFi-CS	0.491	0.618	0.49
SpecFi-CS-a	0.495	0.627	0.486

Table 3: Evaluation on CO. Qwen3-E-4B=Qwen3-Embedding-4B

Setup/Model	MAP	NDCG @10	Avg. R-Prec.
BM25	0.311	0.378	0.219
Qwen3-E-4B	0.502	0.598	0.374
SpecFi-DR	0.443	0.621	0.370
SpecFi-DR-a	0.386	0.536	0.275
SpecFi-CS	0.458	0.626	0.372
SpecFi-CS-a	0.471	0.640	0.386

Table 4: Evaluation on PN

Component Analysis To further analyze the influence of the components of our system, we run different ablation studies documented in Table 5. Here, our main goal is to provide comparison between the proposed SpecFi setups and results from sparse and dense retrieval only, since these performance metrics are also the base for our statistical analysis of correlation. We also include results for

Setup/Model	MAP	NDCG @10	Avg. R-Prec.	s / narrative
NodeRAG only	0.259	0.506	0.323	1.931
BM25	0.080	0.125	0.119	0.011
thenlper/gte-large	0.215	0.394	0.272	2.092
OpenAI-E	0.262	0.507	0.323	0.452
Qwen3-E-4B	0.299	0.523	0.352	6.645
Qwen3-E-4B-p	0.316	0.536	0.370	6.593
CS-direct	0.357	0.536	0.370	1.300
SpecFi-CS-a	0.494	0.726	0.487	14.80

Table 5: Retrieval performance of individual pipeline components on CARDS, serving as baselines for the statistical analysis in §5. Models: OpenAI-E=text-embedding-3-large, Qwen3-E-4B=Qwen3-Embedding-4b. Runtimes were measured on a system with a H100 GPU.

NodeRAG only, where we patched the framework to retrieve the full list of top k results directly. To isolate the contribution of hypothetical document generation, we evaluate CS-direct, which uses the community summary as a direct query expansion without any generation step. CS-direct achieves a MAP of 0.357, above the dense baseline (0.299) but substantially below SpecFi-CS-a (0.494), indicating that the community summaries provide modest retrieval benefit as query expansions but that the majority of SpecFi-CS’s performance gain is attributable to the speculative generation step.

Refusal and Abliteration Analysis. To assess whether the Gemma models refused to generate disinformation-aligned texts, we scanned all generated hypothetical documents ($n=170$ per model) for refusal indicators including direct refusals, role-breaking statements, and safety-related language. Neither the ablated (G-a) nor the non-ablated (G) variant produced any refusals (0% refusal rate). However, the two models differ in output length: G produces longer texts in 110 out of 170 paired generations (mean 48.0 vs. 41.2 words). Since HyDE retrieval relies on cosine similarity between generated and corpus texts in embedding space, we hypothesize that the ablated model’s more concise outputs favor direct claims over verbose qualifications and yield embeddings closer to the shorter, assertive texts typical of disinformation samples in CARDS, consistent with the performance advantage of G-a over G observed across all few-shot configurations in Table 2.

Number of Hypothetical Documents We ablated $n \in \{1, 5, 10, 20\}$ for SpecFi-CS-a on CARDS to evaluate the influence on retrieval performance. MAP increases from 0.438 ($n=1$) to 0.484 ($n=5$) and plateaus at 0.494 ($n=10$) and 0.491 ($n=20$),

Setup	Dataset	Narrative Distinct.	Narrative Variance
BM25	CARDS	-0.240	-0.525*
	CO	-0.357	-0.071
	PN	0.369*	0.319*
QWEN-E-4B	CARDS	-0.066	-0.556*
	CO	-0.679	0.000
	PN	0.197	0.151
SpecFi-DR-a	CARDS	0.147	-0.578*
	CO	-0.964	0.214
	PN	-0.016	0.476**
SpecFi-CS-a	CARDS	0.282	-0.324
	CO	-0.786**	-0.286
	PN	-0.041	0.249

Table 6: Spearman’s ρ between MAP and narrative metrics. FDR-corrected significance: * $p < 0.05$, ** $p < 0.01$.

while runtime scales approximately linearly in n , making $n=10$ a practical tradeoff between retrieval performance and computational cost.

Exploratory Transfer to CO and PN-CC. We further compare performance metrics (Table 3 and Table 4) and possible correlations (Table 6) on two other datasets.

5. Statistical Analysis

For all tests, we compute Spearman’s ρ with FDR correction following the Benjamini–Hochberg procedure. Table 6 reports correlations between MAP and both narrative metrics across datasets. We treat these metrics as competing operationalizations of a narrative’s interpretive space and ask which, if any, is associated with retrieval difficulty.

On CARDS, narrative variance shows consistent negative correlations with MAP across all four systems, reaching significance for BM25, QWEN-E-4B, and SpecFi-DR-a (Table 6). Narrative distinctness does not reach significance on CARDS in the uncontrolled analysis, suggesting that retrieval difficulty is driven by the overall embedding spread within a narrative rather than by inter-narrative separation (D_i). On CO, correlations should be interpreted with caution given the limited number of narratives ($K = 7$); the only significant result is a negative correlation between narrative distinctness and SpecFi-CS-a ($\rho = -0.786$, $p < 0.01$). On PN, the positive correlations between MAP and narrative variance (e.g., BM25: $\rho = +0.319$; SpecFi-DR-a: $\rho = +0.476$) run opposite to the pattern observed on CARDS. We attribute this reversal to two properties of the PN dataset: per-narrative sample sizes are very small (mean $m_i = 2$), making variance estimates unreliable, and the multi-label annotation structure conflates intra-narrative spread with cross-

narrative overlap. We therefore restrict our narrative metric analysis to CARDS, where per-narrative sample sizes ($m_i \in [21, 225]$) support reliable estimation. Leave-one-out analysis confirms that no single narrative, including those with the smallest sample sizes, drives the observed correlations on CARDS.

<i>Original correlations</i>		
Setup	D_i	V_i
BM25	-0.240 (.530)	-0.525 (.123)
QWEN-E-4B	-0.066 (.874)	-0.556 (.122)
SpecFi-DR-a	0.147 (.704)	-0.578 (.122)
SpecFi-CS-a	0.282 (.468)	-0.324 (.468)
<i>Partial correlations (controlling for m_i)</i>		
Setup	D_i	V_i
BM25	-0.029 (.978)	-0.772 (.003)
QWEN-E-4B	-0.007 (.978)	-0.750 (.003)
SpecFi-DR-a	0.125 (.759)	-0.581 (.058)
SpecFi-CS-a	0.387 (.249)	-0.333 (.304)

Table 7: Spearman’s ρ between MAP and narrative metrics on CARDS. FDR-corrected p -values; **bold** $p < 0.05$.

Controlling for Sample Size The number of texts per narrative (m_i) varies from 21 to 225 on CARDS and may itself correlate with both MAP and narrative metrics. We compute partial Spearman correlations by residualizing both MAP and each metric against m_i via linear regression. Table 7 reports results for both metrics; Figure 2 visualizes the relationship for narrative variance. Here, the partial correlations strengthen relative to the uncontrolled analysis: BM25 moves from $\rho = -0.525$ to $\rho = -0.772$ and QWEN-E-4B from $\rho = -0.556$ to $\rho = -0.750$, both significant after FDR correction ($p = 0.003$). SpecFi-DR-a shows a borderline effect ($\rho = -0.581$, $p_{\text{FDR}} = 0.058$; raw $p = 0.014$), significant in all 17 LOO iterations but not after FDR correction; while SpecFi-CS-a remains non-significant ($\rho = -0.333$, $p_{\text{FDR}} = 0.304$). Two-tailed permutation tests (10,000 iterations) confirm these results ($p_{\text{perm}} = 0.0007, 0.0011, 0.014$, and 0.196 for BM25, QWEN-E-4B, SpecFi-DR-a, and SpecFi-CS-a, respectively).

Narrative distinctness remains non-significant throughout. Together, these results indicate that between the two embedding-based narrative metrics, it is the overall intra-narrative spread (V_i), not inter-narrative separation (D_i), that correlates with retrieval difficulty. This is consistent with the interpretation that standard retrieval degrades when a narrative manifests through many diverse framings, rather than when it is merely close to neighboring narratives in embedding space. A median split on V_i (Figure 3) quantifies this effect: BM25 loses 63.4% of its MAP when moving from low- to high-

variance narratives, QWEN-E-4B loses 51.8%, and SpecFi-DR-a loses 41.3%. SpecFi-CS-a shows the smallest degradation (32.7%) while maintaining the highest absolute MAP in both groups. Leave-one-out analysis confirms stability: partial correlations remain significant in all 17 iterations for BM25, QWEN-E-4B, and SpecFi-DR-a, with no single narrative acting as a leverage point (BM25 LOO range: $\rho \in [-0.83, -0.73]$).

6. Discussion

Analysis of the community summaries retrieved for each CARDS narrative reveals a key insight: the community summaries are generated without access to narrative labels, NodeRAG constructs its knowledge graph and applies Leiden community detection exclusively on the textual content of the training corpus. Of the 17 CARDS narratives, 11 receive summaries that align with the taxonomy label at least at the super-claim level, 2 are collapsed with a sibling sub-narrative, and 4 exhibit drift or incoherence (full mapping is provided in Table 9 in the appendix). For instance, the summary retrieved for narrative 1_2 (“heading into ice age / global cooling”) independently arrives at “the Earth may be entering a cooling cycle,” and narrative 5_1 (“science is uncertain / unreliable”) yields “skepticism about the reliability of climate models.” This convergence suggests that the CARDS narrative taxonomy reflects genuine topical structure in the disinformation corpus rather than predefined classification, and that graph-based community detection can surface this structure from unlabeled text; a property with application for monitoring emerging narratives that lack predefined labels.

Instances where the summaries fail reveal diagnostic patterns that help explain the system’s behavior. The collapse pattern (where sub-narratives such as 4_1/4_2 or 3_2/3_3 receive identical summaries) correlates directly with low per-narrative AP for SpecFi-CS and identifies Leiden resolution as a tunable parameter. The drift pattern reveals a subtler problem: narrative 4_4 (“clean energy won’t work”) receives a summary that argues *for* technological solutions, inverting the narrative’s stance. This polarity inversion arises because community detection clusters texts by topic co-occurrence, which does not inherently distinguish argumentative direction. Texts criticizing and texts promoting renewable energy share entities and relationships (solar panels, wind turbines, efficiency, cost), so Leiden groups them together, and the LLM’s summary reflects the majority framing. This failure mode suggests that strategies like stance-aware community detection could address a class of errors that finer resolution alone would not resolve.

These failure patterns are consistent with SpecFi-

Partial Correlation: MAP vs. Narrative Variance on CARDS

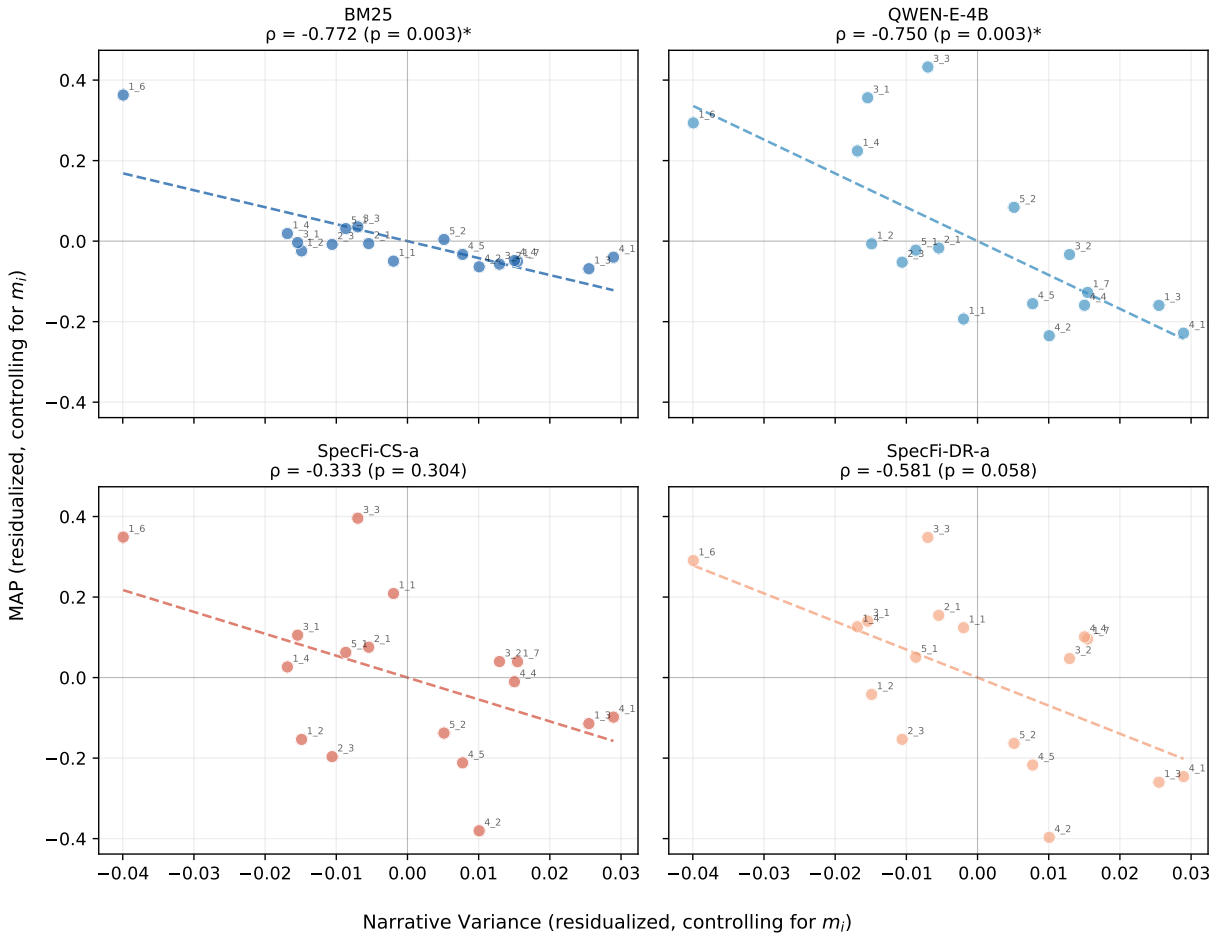


Figure 2: Partial correlation between MAP and narrative variance on CARDS, controlling for m_i . Each point represents one narrative. BM25 and QWEN-E-4B show steep negative slopes; SpecFi-CS-a shows no significant trend. All p -values are FDR-corrected (Benjamini–Hochberg across all tests in Table 7).

CS-a showing the lowest inter-system correlation with BM25 on CARDS ($\rho = 0.365$, $p = 0.249$), indicating that the community-summary-based system retrieves narratives through a qualitatively different mechanism than dense or lexical retrieval, producing complementary errors. Where community summaries converge on the correct narrative premise, SpecFi-CS generates hypotheticals that span the narrative’s interpretive range—as illustrated by narrative 3_3 (“CO₂ is plant food”), where the abstract summary enables generation of diverse hypothetical documents covering CO₂ fertilization, agricultural productivity, and pollutant classification arguments, rather than anchoring on a single text’s framing. Where summaries collapse or drift, the generated hypotheticals lose discriminative power or target the wrong stance entirely.

7. Future Work

While this study focuses on climate change denial, the SpecFi framework is domain-agnostic. Applying it to other narrative datasets (such as European disinformation narratives (Sosnowski et al., 2024), COVID-19 conspiracy narratives (Heinrich et al., 2024), or propaganda taxonomies (Solopova et al., 2023; Sahitaj et al., 2025)) would test the generalizability of both the retrieval approach and the narrative variance metric. On the retrieval side, the final step still relies on dense cosine similarity. Following Hatzel and Biemann (2024) and Akter and Santu (2024), more interpretive similarity measures that incorporate narrative features such as actors, localities, and argumentative structure could be explored. Similarly, aligning the graph representation more closely with narrative systems (Hellman, 2024) could improve both community summary quality and retrieval performance.

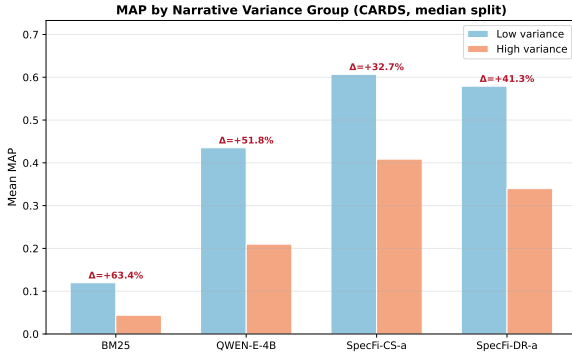


Figure 3: Mean MAP for narratives above and below median V_i on CARDS. BM25 loses 63.4% of its MAP on high-variance narratives; SpecFi-CS-a loses 32.7%.

8. Conclusion

In this study, we re-framed climate disinformation detection as a narrative retrieval task and introduced SpecFi, a speculative-document generation framework that bridges abstract narrative descriptions and their diverse textual realizations. Across three datasets, SpecFi, and especially the community-summary variant, improves robustness compared to sparse and dense baselines, remaining stable even for high-variance narratives. Our analysis further shows that narrative variance correlates with retrieval difficulty for standard baselines, while graph-derived community summaries can recover narrative structure from unlabeled data. Together, these results highlight narrative retrieval as a flexible approach for tracking evolving disinformation narratives beyond fixed taxonomies.

Limitations

While we were able to provide a version of SpecFi-DR which only relies on open source models to ensure reproducibility, the SpecFi-CS setups include one reliance on OpenAI models within NodeRAG. Recent studies have shown that OpenAI models still outperform on structured output generation (Geng et al., 2025), which is an essential step during graph construction. For this reason and due to NodeRAG’s own recommendation³, we used the proprietary model here. An additional factor that could affect our results: the CARDS dataset is from 2021, making it likely to be part of the training data of LLMs. While this does not necessarily relate to our specific usage of this dataset, it is still possible that there is an influence on the generation

³https://terry-xu-666.github.io/NodeRAG_web/blog/2025/03/16/structure-output/

of hypothetical documents as well as community summaries. However, our results of the zero shot variants in Table 2 indicate that none of our tested LLMs is capable of generating representative hypotheticals without examples and only based on the narrative by itself, but an influence in some kind of capacity cannot be ruled out. Our evaluation relies on automatic retrieval metrics derived from existing narrative annotations; human evaluation of narrative alignment quality remains for future work. Similarly, the convergence analysis between community summaries and expert-crafted taxonomies (Section 6) is based on qualitative judgment. We provide a systematic mapping of all 17 narratives to pattern categories in Table 9 in the appendix for verification, but a more rigorous evaluation with independent annotators would strengthen this claim.

Ethical Considerations

Recent work has shown that current LLMs can generate convincing disinformation following predefined narratives (Vykopal et al., 2024) and that personalization requests can bypass safety filters (Zugecova et al., 2025), highlighting the dual-use risk of methods built around disinformation generation, including ours. Although our method is targeted towards counter-disinformation efforts, it could also encourage further fine-tuning of LLMs to improve generating disinformation. Within this study, we only use models already available on huggingface. This point needs to be taken into account further in future work, like the question whether models fine-tuned for generating disinformation should be released publicly and if so, how the release can be controlled while also indicating ethical considerations, e.g., in model cards.

Acknowledgments

The work on this paper is performed in the scope of the projects “VeraXtract” (16IS24066) and “news-polygraph” (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

9. Bibliographical References

- Mousumi Akter and Shubhra Kanti Karmaker Santu. 2024. [Fans: a facet-based narrative similarity metric.](#)
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global:](#)

- A graph rag approach to query-focused summarization.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.
- Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. [Jsonschemabench: A rigorous benchmark of structured outputs for language models](#).
- Fatima Haouari, Carolina Scarton, Nicolò Faggiani, Nikolaos Nikolaidis, Bonka Kotseva, Ibrahim Abu Farha, Jens Linge, and Kalina Bontcheva. 2025. [UKElectionNarratives: A Dataset of Misleading Narratives Surrounding Recent UK General Elections](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19:2477–2495.
- Hans Ole Hatzel and Chris Biemann. 2024. [Story embeddings — narrative-focused representations of fictional stories](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.
- Philipp Heinrich, Andreas Blombach, Bao Minh Doan Dang, Leonardo Zilio, Linda Havenstein, Nathan Dykes, Stephanie Evert, and Fabian Schäfer. 2024. [Automatic identification of COVID-19-related conspiracy narratives in German telegram channels and chats](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1932–1943, Torino, Italia. ELRA and ICCL.
- Maria Hellman. 2024. [Narrative Analysis and Framing Analysis of Disinformation](#), pages 101–121. Springer Nature Switzerland, Cham.
- Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2025. [A discourse analysis framework for legislative and social media debates](#). In *Proceedings of the 17th ACM Web Science Conference 2025, Websci '25*, page 199–209, New York, NY, USA. Association for Computing Machinery.
- Bonka Kotseva, Irene Vianini, Nikolaos Nikolaidis, Nicolò Faggiani, Kristina Potapova, Caroline Gasparro, Yaniv Steiner, Jessica Scornavacche, Guillaume Jacquet, Vlad Dragu, Leonida della Rocca, Stefano Bucci, Aldo Podavini, Marco Verile, Charles Macmillan, and Jens P. Linge. 2023. [Trend analysis of COVID-19 mis/disinformation narratives—A 3-year study](#). *PLOS ONE*, 18(11):e0291423. Publisher: Public Library of Science.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#).
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- OpenAI. 2024. [Gpt-4o system card](#).
- Andrew Piper, Richard Jean So, and David Bamman. 2021. [Narrative theory for computational narrative understanding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hanna-Riikka Roine. 2020. On speculation as a strategy. *Fafnir – Nordic Journal of Science Fiction and Fantasy Research*, 7(2):8–15.
- Ariana Sahitaj, Premtim Sahitaj, Veronika Solopova, Jiaao Li, Sebastian Möller, and Vera Schmitt. 2025. [Hybrid annotation for propaganda detection: Integrating LLM pre-annotations with human intelligence](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 215–228, Vienna, Austria. Association for Computational Linguistics.
- Yanzhen Shen, Sihao Chen, Xueqiang Xu, Yunyi Zhang, Chaitanya Malaviya, and Dan Roth. 2025. [Logicol: Logically-informed contrastive learning for set-based dense retrieval](#).
- Iknoor Singh. 2024. [Detecting and Tracking the Spread of Debunked Narratives Across Languages](#). Phd thesis, University of Sheffield, Sheffield, UK. Supervisors: Carolina Scarton and Kalina Bontcheva.
- Iknoor Singh, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2024. [Breaking language barriers with mmtweets: Advancing cross-lingual debunked narrative retrieval for fact-checking](#).
- Veronika Solopova, Christoph Benz Müller, and Tim Landgraf. 2023. [The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective](#). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

- Witold Sosnowski, Arkadiusz Modzelewski, Kinga Skorupska, Jahna Otterbacher, and Adam Wierzbicki. 2024. [EU DisinfoTest: a benchmark for evaluating language models' ability to detect disinformation narratives](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14702–14723, Miami, Florida, USA. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. 2025. [Gemma 3 technical report](#).
- V. A. Traag, L. Waltman, and N. J. van Eck. 2019. [From Louvain to Leiden: guaranteeing well-connected communities](#). *Scientific Reports*, 9(1):5233.
- Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. [Disinformation capabilities of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14847, Bangkok, Thailand. Association for Computational Linguistics.
- Tianyang Xu, Haojie Zheng, Chengze Li, Haoxiang Chen, Yixin Liu, Ruoxi Chen, and Lichao Sun. 2025. [Noderag: Structuring graph-based rag with heterogeneous nodes](#).
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. 2025. [Qwen3 technical report](#).
- Yan Zhang, Zhaopeng Feng, Zhiyang Teng, Zuozhu Liu, and Haizhou Li. 2023. [How well do text embedding models understand syntax?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9717–9728, Singapore. Association for Computational Linguistics.
- Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. [Are NLP models good at tracing thoughts: An overview of narrative understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.
- Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. [Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

10. Language Resource References

- Coan, Travis G. and Boussalis, Constantine and Cook, John and Nanko, Mirjam O. 2021. *CARDS: Computer-Assisted Recognition of Denial and Skepticism – Climate Change Contrarian Claims Dataset*. distributed with the article in Scientific Reports. PID <https://doi.org/10.1038/s41598-021-01714-4>. Dataset of climate change denial claims organized under a two-level taxonomy of super-claims and sub-claims.
- Nikolaidis, Nikolaos and Stefanovitch, Nicolas and Silvano, Purificação and Dimitrov, Dimitar Iliyanov and Yangarber, Roman and Guimarães, Nuno and Sartori, Elisa and Androutsopoulos, Ion and Nakov, Preslav and Da San Martino, Giovanni and Piskorski, Jakub. 2025. *PolyNarrative: A Multilingual, Multilabel, Multi-domain Dataset for Narrative Extraction from News Articles*. Association for Computational Linguistics. PID <https://doi.org/10.18653/v1/2025.acl-long.1513>. Multilingual dataset of news articles annotated with fine-grained narrative labels across multiple topics.
- Rowlands, Harri and Morio, Gaku and Tanner, Dylan and Manning, Christopher. 2024. *Climate Obstruction in Social Media Advertising Dataset*. Association for Computational Linguistics. PID <https://doi.org/10.18653/v1/2024.findings-acl.330>. Dataset of social media advertisements from oil and gas companies classified under obstruction narratives.

A. Appendix

A.1. Evaluation Details

A.1.1. Experimental Setup

All experiments were run on a system with an NVIDIA H100 GPU. The runtimes for setups based on OpenAI models reflect the inference time behind the OpenAI API.

A.1.2. Performance per Narrative on CARDS and Statistical Analysis of Performance Correlation

As documented in Figure 4, although the retrieval performance differs per narrative when compared over the whole CARDS dataset, there are also similarities between the results across setups. For example, the two best results for all setups are the spikes at the narrative ids 1_6 and 3_3, while 4_1 and 4_2 are some of the lowest scores for all 4 setups, notably also including SpecFi-CS.

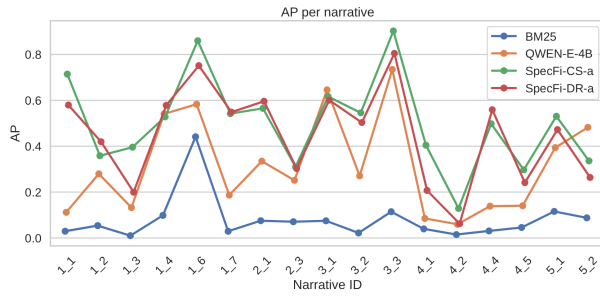


Figure 4: AP results per narrative on the CARDS dataset

Inspecting the community summaries retrieved as few-shot examples for each narrative reveals three distinct patterns: convergence, collapse, and drift (see Appendix A.2 for the full mapping).

Convergence. For the majority of narratives, the retrieved community summary closely mirrors the expert-crafted narrative label from the CARDS taxonomy, despite the community detection operating without access to any labels. For example, the summary retrieved for narrative 1_2 (“heading into ice age / global cooling”) states that “the Earth may be entering a cooling cycle,” and the summary for 5_1 (“science is uncertain / unreliable”) expresses “skepticism about the reliability of climate models.” Similar convergence is observed for narratives 1_6, 2_1, 3_1, and 5_2. We verified in the NodeRAG source code that neither filenames nor any external metadata enter the graph construction or summary generation pipeline: the LLM operates exclusively on text content extracted from the corpus.⁴ The convergence therefore reflects genuine bottom-up re-derivation of narrative structure from textual co-occurrence patterns in the knowledge graph.

Collapse. Where the Leiden resolution is too coarse, semantically adjacent sub-narratives merge into a single community. Narratives 4_1 (“climate policies are harmful”), 4_2 (“policies are ineffective”), and 4_3 (“too difficult to solve”) all receive an identical candidate summary (“Current climate policies are criticized for being ineffective, as they rely on unrealistic targets and fail to consider political and technological realities, leading to significant market failures.”), collapsing distinct argumentative strategies into a single description.

⁴Specifically, `CommunitySummary.get_normal_query()` in the NodeRAG codebase aggregates the `context` field of semantic unit and attribute nodes within each Leiden partition; input filenames are stored only in a separate document tracking table and never appear in any LLM prompt.

Similarly, narratives 3_2 and 3_3 share a summary about CO₂ increases being “harmless or even beneficial.” These narratives show correspondingly low AP for SpecFi-CS-a, suggesting that the system’s few-shot examples lack the specificity needed to generate discriminative hypotheticals when sub-narratives share thematic structure. This points to community detection granularity as a key parameter for future optimization: finer-grained communities could preserve distinctions that the current Leiden resolution merges.

Drift. A third failure mode occurs when the community captures the *topic* but not the *stance*. The summary retrieved for narrative 4_4 (“clean energy technology / biofuels won’t work”) instead describes “advancements in technology” that “can provide innovative solutions”—effectively arguing *for* clean energy rather than against it. This polarity inversion likely arises because the community was dominated by texts *discussing* renewable energy technology, and the LLM’s summarization defaulted to the majority framing within the cluster. As a result, the generated hypotheticals are semantically opposed to the target narrative, representing a fundamentally different failure from collapse: where collapse loses granularity, drift inverts argumentative direction.

In Table 8, we evaluate if the performance of the considered systems does indeed correlate. Several statistically significant correlations can be reported, especially within the results on CARDS and PN. For example, the comparison between BM25 and QWEN3-E-4B indicates the highest correlation with a rho value of 0.824 and a FDR-corrected p-value of 0.000 on CARDS. Both SpecFi variants behave more independently, especially in regard to the SpecFi-CS-a setup which, for example, yields the lowest rho values when compared to BM25 results with a p-value of 0.249 on CARDS and thus with the highest value above the 0.05 significance threshold.

A.2. Community Summaries

A.3. Prompts Collection

A.3.1. Embedding Models

The default model instruction prompt is:

```
Instruct: Given a web search query, retrieve relevant passages that answer the query
Query: {query}
```

The following prompt was used to enhance the retrieval results:

```
Instruct: Given a narrative description as a query, retrieve passages that serve this narrative; can be entailed from the
```

#	Metric	Setup	QWEN				
			BM25	-E-4B	SpecFi-DR-a	SpecFi-CS-a	
1	RHO	BM25	1.000	0.824	0.544	0.365	
		QWEN-E-4B	0.824	1.000	0.679	0.520	
		SpecFi-DR-a	0.544	0.679	1.000	0.892	
		SpecFi-CS-a	0.365	0.520	0.892	1.000	
		FDR-p	BM25	0.000	0.000	0.060	0.249
	QWEN-E-4B	0.000	0.000	0.009	0.065		
	SpecFi-DR-a	0.060	0.009	0.000	0.000		
	SpecFi-CS-a	0.249	0.065	0.000	0.000		
	2	RHO	BM25	1.000	0.786	0.321	-0.071
			QWEN-E-4B	0.786	1.000	0.536	0.321
SpecFi-DR-a			0.321	0.536	1.000	0.750	
SpecFi-CS-a			-0.071	0.321	0.750	1.000	
FDR-p			BM25	0.000	0.121	0.536	0.879
QWEN-E-4B		0.121	0.000	0.359	0.536		
SpecFi-DR-a		0.536	0.359	0.000	0.130		
SpecFi-CS-a		0.879	0.536	0.130	0.000		
3		RHO	BM25	1.000	0.657	0.489	0.366
			QWEN-E-4B	0.657	1.000	0.436	0.353
	SpecFi-DR-a		0.489	0.436	1.000	0.558	
	SpecFi-CS-a		0.366	0.353	0.558	1.000	
	FDR-p		BM25	0.000	0.000	0.001	0.014
	QWEN-E-4B	0.000	0.000	0.004	0.016		
	SpecFi-DR-a	0.001	0.004	0.000	0.000		
	SpecFi-CS-a	0.014	0.016	0.000	0.000		

Table 8: Spearman’s rho and FDR-corrected p-values for correlation analysis of performances on the MAP metric. #1 documents the results on CARDS, #2 on CO and #3 on PN.

```
narrative; can be aligned logically with the
narrative
Query: {query}
```

Within our HyDE-based setup, we use the following instruction prompt for embedding models:

```
Instruct: Given a text as a query retrieve
relevant passages that align with narratives
similar to the query
Query: {query}
```

A.3.2. Generative Models

Our system prompt:

```
You are a disinformation investigator. Your first
step is to generate short disinformation
texts that sound like actual ones. You get a
disinformation narrative and return a
disinformation text that aligns with that
narrative. Return only 1 single text!
```

Each system prompt was followed by a user prompt:

```
You are a disinformation investigator. Given a
disinformation narrative, generate a short,
realistic text (such as a news excerpt, blog
post, or social media post) that supports or
aligns with that narrative. The text should
sound plausible and could be found in the
wild.

Here are some examples: {examples}

Narrative: {query}
Text:
```

A.4. Narrative Taxonomies Transformed to Search Queries

All datasets included predefined narrative taxonomies. We constructed narrative descriptions from the narrative labels with the following strategies:

A.4.1. CARDS

The queries were constructed by following the CARDS taxonomy and the corresponding id labeling with the dataset and concatenating “claim” (=subnarrative) and “superclaim” (=narrative) descriptions from the code book.

```
"1_1": "Global warming is not happening.
Ice/permafrost/snow cover isn't melting",
"1_2": "Global warming is not happening. We're
heading into an ice age/global cooling",
"1_3": "Global warming is not happening. Weather
is cold/snowing",
...
"5_3": "Climate movement/science is unreliable.
Climate change (science or policy) is a
conspiracy (deception)",
```

Resulting in predefined 27 narratives in the taxonomy, out of which 17 can be found in the dataset.

A.4.2. Climate Obstruction

Constructed by using the narrative ids from the dataset and descriptions of the narratives provided in the supplemental material of the paper.

```
"CA": "Community & Resilience. Emphasizes how the
oil and gas sector contributes to local and
national economies through tax revenues,
charitable efforts, and support for local
businesses",
"CB": "Community & Resilience. Focuses on the
creation and sustainability of jobs by the
oil and gas industry."
...
"SA": "Patriotic Energy mix. Stresses how domestic
oil and gas production benefits the nation,
including energy independence, energy
leadership, and the idea of supporting
American energy"
```

Resulting in 7 narratives, out of which all can be found in the dataset.

A.4.3. PolyNarrative

Similar to CARDS, the queries were constructed by following the PolyNarrative (PN) taxonomy and the corresponding id labeling with the dataset.

```
"1_1": "Blaming the war on others rather than the
invader: Ukraine is the aggressor",
"1_2": "Blaming the war on others rather than the
invader: The West are the aggressors",
...
"21_2": "Green policies are geopolitical
instruments: Green activities are a form of
neo-colonialism"
```

Resulting in predefined 88 narratives in the taxonomy, out of which 51 can be found in the dataset.

Table 9: CARDS narrative taxonomy with community summaries generated within the NodeRAG framework. Pattern categories: *convergence* (summary aligns with taxonomy label), *partial* (aligns at super-claim level), *collapse* (identical summary shared with sibling sub-narrative), *drift* (correct topic, wrong stance or focus), *incoherent* (summary unrelated to narrative).

ID	Narrative Label	Pattern	Community Summary
1: Global warming is not happening			
1_1	Ice/permafrost/snow cover isn't melting	partial	The text explores the concept of anthropogenic global warming as a myth, questioning its validity and the narratives surrounding it.
1_2	Heading into ice age/global cooling	converg.	There are emerging voices cautioning against the narrative of catastrophic global warming, suggesting that the Earth may be entering a cooling cycle.
1_3	Weather is cold/snowing	drift	Severe weather events, such as unexpected snowfall, significantly affect city operations and highlight the need for preparedness in urban planning.
1_4	Climate hasn't warmed over the last decade(s)	partial	Maps generated by climate models, used by the IPCC, are criticized as 'fantasy maps' that do not accurately reflect Earth's climate history or current state.
1_6	Sea level rise is exaggerated/not accelerating	converg.	Recent studies indicate that the rate of sea level rise has remained consistent, contradicting some climate model predictions, which raises questions about their reliability.
1_7	Extreme weather isn't increasing/has happened before	converg.	The relationship between climate change and extreme weather events remains contentious. While some studies suggest that rising temperatures may lead to more severe weather patterns, others argue that evidence does not support a significant increase in the incidence or severity of such events.
2: Human greenhouse gases are not causing climate change			
2_1	It's natural cycles/variation	converg.	The debate over climate change often centers on the relative contributions of natural variability versus human-induced factors. While greenhouse gas emissions are acknowledged, many scientists emphasize the significant role of natural processes in shaping climate.
2_3	No evidence for greenhouse effect/CO ₂ driving climate change	drift	Climate change poses significant challenges to agriculture, with erratic weather patterns threatening crop yields. However, rising CO ₂ levels may enhance plant growth, presenting a complex scenario where adaptation strategies are essential.
3: Climate impacts/global warming is beneficial/not bad			
3_1	Climate sensitivity is low/negative feedbacks	converg.	Recent studies suggest that the negative impacts of warming may not be as severe as previously believed, indicating that CO ₂ climate sensitivity is significantly lower than earlier estimates.
3_2	Species/plants/reefs benefiting from climate change	converg.	Some argue that the effects of CO ₂ increases and slight global warming may be harmless or even beneficial, challenging alarmist narratives about climate change.
3_3	CO ₂ is beneficial/plant food	collapse	(Same as 3_2)
4: Climate solutions won't work			
4_1	Climate policies are harmful	converg.	Current climate policies are criticized for being ineffective, as they rely on unrealistic targets and fail to consider political and technological realities, leading to significant market failures.
4_2	Climate policies are ineffective/flawed	collapse	(Same as 4_1)
4_4	Clean energy/biofuels won't work	drift	The text explores how advancements in technology can provide innovative solutions to combat climate change, including renewable energy sources and carbon capture methods.
4_5	People need energy (fossil fuels/nuclear)	incoherent	The presence of the Wolverine at the convention stage highlights how external factors can influence the performance and effectiveness of energy panels, suggesting a need for adaptive strategies in energy management.
5: Climate movement/science is unreliable			
5_1	Science is uncertain/unsound/unreliable	converg.	Many scientists express skepticism about the reliability of climate models, which have been criticized for failing to accurately predict temperature changes and for being overly reliant on theoretical calculations.
5_2	Movement is alarmist/political/biased	converg.	The authors argue that the climate alarmism movement is losing credibility, with outdated predictions and ideological biases undermining its claims.

Unsupervised GRI-TCFD Alignment with LLM-Assisted Validation for Climate Disclosure and Greenwashing Risk Analysis

Seyed Alireza Mousavian Anaraki, Danilo Croce,
Roberta Costa, Luigi Tiburzi, Armando Calabrese and Roberto Basili

Department of Enterprise Engineering
University of Rome Tor Vergata, Via del Politecnico 1, 00133, Rome, Italy
seyedalireza.mousaviananaraki@students.uniroma2.eu
{croce, basili}@info.uniroma2.it
{roberta.costa, luigi.tiburzi}@uniroma2.it
calabrese@dii.uniroma2.it

Abstract

Climate-related corporate disclosures play a central role in sustainable finance and regulatory supervision, but remain difficult to analyze due to their length, unstructured format, and strategic language. While existing NLP approaches have been applied to ESG scoring and greenwashing detection, most operate at the document level and lack explicit alignment with formal reporting standards. We propose a scalable paragraph-level framework for aligning sustainability disclosures with the Global Reporting Initiative (GRI) indicators and the Task Force on Climate-related Financial Disclosures (TCFD) pillars. Our approach combines weak supervision, climate-focused GRI-TCFD mapping, embedding-based semantic similarity, and LLM validation for climate detection. In parallel, we introduce a paragraph-level greenwashing proxy based on commitment intensity, claim specificity, and sentiment polarity. This proxy complements regulatory alignment by capturing linguistic signals associated with potentially symbolic climate communication. The resulting augmented data are used to fine-tune ClimateBERT models in both single-task and multi-task settings. Experimental results show that weakly supervised dataset augmentation improves robustness and generalization compared to purely manual training, with further gains in the multi-task configuration. By integrating regulatory semantics, domain-adapted language models, and scalable annotation strategies, this study advances standard-aligned climate disclosure analysis and provides tools directly relevant to climate-related financial risk assessment.

Keywords: Sustainability reporting, GRI, TCFD, Greenwashing Risk Proxy, ClimateBERT

1. Introduction

Climate-related disclosure has become a central pillar of sustainable finance and regulatory oversight. Institutional investors, banks, and supervisory authorities increasingly rely on structured reporting frameworks to assess exposure to transition and physical risks. Among the most widely adopted standards, the Global Reporting Initiative (GRI) provides granular sustainability indicators, while the Task Force on Climate-related Financial Disclosures (TCFD) defines a financial-material framework organized around Governance, Strategy, Risk Management, and Metrics and Targets. These frameworks directly inform regulatory supervision, capital allocation, and climate risk modeling (Bingler et al., 2022).

At the same time, sustainability reporting has expanded dramatically in volume and complexity. Reports are typically long, unstructured PDF documents written in strategic corporate language, with climate-related information scattered across sections (Anaraki et al., 2025). This makes paragraph-level interpretation difficult. Moreover, a growing body of literature reports on selective disclosure and “greenwashing” practices, in which nar-

ative emphasis may exceed substantive commitment (Janik and Ryszko, 2025). Empirical evidence shows that discursive Environmental, Social, and Governance (ESG) emphasis may not always translate into proportional strategic or operational changes (Bingler et al., 2024; Wood et al., 2025).

Recent advances in natural language processing (NLP) and large language models (LLMs) have enabled automated extraction, classification, and verification of sustainability-related information (Moodaley and Telukdarie, 2023a). Transformer-based and LLM-based approaches have been used to detect green practices in social media (Glazkova and Zakhrova, 2025), assess ESG commitment in financial documents (Wood et al., 2025), identify green claims and greenwashing (Moodaley and Telukdarie, 2023b), and perform climate-specific classification through domain-adapted models such as ClimateBERT (Webersinke et al., 2022). These tools support scalable and systematic analysis of corporate disclosures across ESG dimensions (Zou et al., 2025; Kazakov et al., 2023), facilitating better alignment with reporting frameworks such as GRI (Ngee et al., 2024; Bronzini et al., 2024), the Sustainable Development Goals (SDGs) (Jakob et al., 2024;

Li and Rockinger, 2024), and TCFD, particularly through domain-adapted BERT variants such as ClimateBERT for greenwashing detection (Bingler et al., 2022). Prior work has also explored weakly supervised and unsupervised methods for automatically aligning report content with multiple sustainability standards, particularly focusing on integrating GRI and SDG (Mousavian Anaraki et al., 2025a,b).

However, existing work largely focuses on document-level scoring, sentiment detection, or general climate relevance, rather than structured paragraph-level alignment with formal regulatory architectures.

We address this gap by proposing a scalable paragraph-level annotation framework that leverages the officially published alignment between GRI indicators and TCFD pillars to guide climate-related text identification. Rather than directly training a regulatory classifier, we use GRI-TCFD alignment as structured weak supervision to construct high-confidence paragraph-level climate annotations. Our approach combines weak supervision from GRI content indices, embedding-based semantic similarity, and LLM-assisted validation. MP-Net (Song et al., 2020) embeddings rank paragraph-to-definition matches across candidate GRI-TCFD pairs, while GPT-OSS (Agarwal et al., 2025) filters noisy assignments through definition-grounded reasoning. This process yields a standard-informed climate detection dataset without requiring manual paragraph-level expert annotation. For instance, the following excerpt from a sustainability report:

“Our goal is to be a leader in ecologically sound production by looking at our own carbon footprint and aiming for the highest possible standards of ecological responsibility among our producers.”

is automatically classified as climate-related and aligned with the corresponding GRI-TCFD categories:

- **GRI 305** (EMISSIONS), specifically **GRI 305-5**: “Reduction of GHG emissions”
- **TCFD** (GOVERNANCE): “Disclose the organization’s governance around climate-related risks and opportunities”
- **Climate-related** (EMISSIONS, GOVERNANCE)

In parallel, we introduce a paragraph-level greenwashing risk proxy derived from commitment intensity, claim specificity, and sentiment polarity. Inspired by (Vinella et al., 2024), we define a greenwashing risk proxy based on three characteristic patterns commonly observed in misleading sustainability disclosures: (1) absence of explicit climate-related commitments and actions, (2) use of non-

specific or vague language, and (3) overly optimistic or promotional sentiment. Each dimension is treated as an indicator, and their combination yields a greenwashing risk proxy, which is discretized into *Low-risk* (one active indicator), and *High-risk* (two or more active indicators) categories. Following this formulation, the paragraph is labeled as:

- **COMMITMENT**: Inactive
- **NON-SPECIFIC**: Active
- **OPTIMISTIC SENTIMENT**: Active

resulting in a HIGH GREENWASHING RISK label. The proxy is computed only on validated climate-related paragraphs and captures linguistic patterns associated with potentially symbolic climate communication. Rather than replacing regulatory alignment, this proxy complements climate relevance by providing an additional supervisory signal grounded in discourse characteristics. The automatically constructed datasets are then used for benchmark augmentation.

We fine-tune ClimateBERT in single-task (climate relevance) and multi-task (climate relevance and greenwashing risk) settings to evaluate whether standard-informed weak supervision improves downstream performance. By coupling regulatory structure with scalable annotation and multi-task learning, our framework reduces reliance on costly manual labeling while maintaining semantic consistency with established reporting standards and improving empirical robustness in downstream classification.

This paper makes three main contributions:

1. We propose a scalable, weakly supervised framework for climate-related paragraph-level alignment of corporate disclosures with GRI and TCFD standards, combining semantic similarity and LLM-based validation;
2. We introduce a transparent, literature-grounded paragraph-level greenwashing risk proxy based on commitment, specificity, and sentiment attributes;
3. We show that automatically generated annotations improve climate detection and greenwashing risk classification through dataset augmentation and multitask learning.

To guide our empirical evaluation, we address the following research questions:

- **RQ1**. Does augmenting benchmark datasets with automatically constructed annotations improve downstream performance?
- **RQ2**. Does joint multi-task learning provide additional gains over separate single-task models?

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed paragraph-level GRI-TCFD alignment framework and the greenwashing risk proxy construction. Section 4 reports the experimental evaluation. Section 5 concludes the paper.

2. Related Work

Natural language processing has been widely applied to the analysis of corporate sustainability disclosures. Prior work employs sentence similarity, sentiment classification, and information extraction techniques to assess environmental risk and ESG communication patterns in unstructured reports (Kang and Kim, 2022; Polignano et al., 2022). These approaches enable large-scale analysis but typically operate at the document level or focus on isolated linguistic signals.

The emergence of LLMs has further expanded capabilities for extracting structured information and detecting misleading or greenwashing claims in sustainability reporting (Moodaley and Telukdarie, 2023a,b). Domain-adapted models such as ClimateBERT (Webersinke et al., 2022) incorporate climate-specific pretraining and have demonstrated strong performance across tasks including climate risk detection (Garrido-Merchán et al., 2026), cheap talk analysis (Bingler et al., 2024), and greenwashing identification (Vinella et al., 2024).

However, progress remains constrained by the scarcity of high-quality paragraph-level annotations. Recent efforts such as Climate-NLI (Yudanto et al., 2024) explore zero- and few-shot classification through natural language inference (Yin et al., 2019), yet performance is sensitive to semantic overlap and task ambiguity.

In parallel, weakly supervised and unsupervised approaches have been proposed to align sustainability disclosures with structured reporting standards, particularly for integrating GRI and SDG frameworks (Mousavian Anaraki et al., 2025a,b). Nevertheless, structured paragraph-level alignment with formal climate-related architectures such as the GRI-TCFD mapping remains underexplored.

Our work builds on these strands by combining structured weak supervision, embedding-based similarity modeling, and LLM-assisted validation to construct a paragraph-level, standard-informed climate dataset. Unlike prior studies that primarily address document-level scoring or standalone classification tasks, we explicitly leverage the official GRI-TCFD alignment to guide annotation and evaluate its impact through benchmark augmentation and multi-task fine-tuning.

3. Standard-Informed Paragraph-Level Annotation Framework

This section presents a scalable, weakly supervised pipeline for constructing paragraph-level annotated datasets for (i) climate relevance detection and (ii) greenwashing risk classification from corporate sustainability reports. The framework follows a sequential two-stage design, where the output of the first stage directly feeds into the second. In the first stage, we construct a climate detection dataset by identifying climate-related paragraphs through GRI-TCFD alignment, embedding-based semantic similarity, and LLM-assisted validation. The objective is to maximize annotation precision while avoiding manual expert labeling, leveraging structured weak supervision from reporting standards. In the second stage, we derive a greenwashing risk dataset by restricting the analysis to the validated climate-related subset and assigning commitment, specificity, and sentiment attributes via few-shot prompting, following (Bingler et al., 2024). These attributes are aggregated into a composite greenwashing risk proxy. By grounding greenwashing assessment in previously validated climate-related paragraphs, the framework ensures conceptual consistency between climate relevance and risk characterization. The automatically constructed datasets are subsequently used for benchmark augmentation and downstream evaluation (Section 4).

3.1. Automatic Climate Paragraph Detection via GRI-TCFD Alignment

Our approach is inspired by prior work on automatic alignment between GRI and SDG standards (Mousavian Anaraki et al., 2025b), which demonstrated the effectiveness of combining weak supervision, semantic similarity, and structured standards for large-scale sustainability annotation. Building upon this foundation, we extend the methodology to the climate reporting domain by integrating the TCFD framework and introducing LLM-assisted validation. The objective is to construct a high-quality paragraph-level dataset D , with $|D| = M$, derived from N different corporate disclosures (paragraphs) (with $M < N$), augmented with standardized climate reporting categories, without requiring manual expert annotation. An overview of the framework is shown in Figure 1. The proposed framework consists of six main stages: (i) paragraph extraction and pre-processing, (ii) weak GRI label initialization from content indices, (iii) climate-focused GRI-TCFD pairing, (iv) embedding-based similarity scoring, (v) indexed and non-indexed label disambiguation, and (vi) LLM-assisted validation.

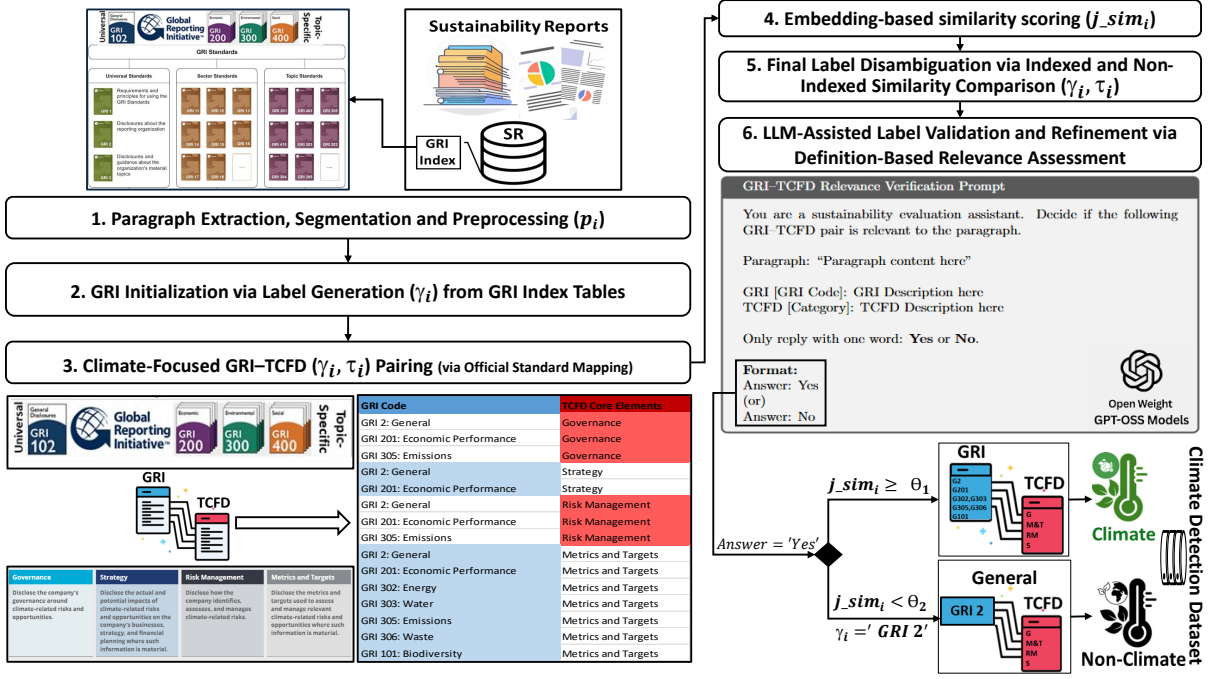


Figure 1: Overview of the proposed GRI-TCFD-based climate-related paragraph detection pipeline. The framework integrates weak supervision, semantic similarity modeling, and LLM-based validation.

The output is a structured climate detection dataset:

$$D = \{(p_i, \gamma_i, \tau_i, d(\gamma_i), d(\tau_i), llm_i, j_sim_i, c_i)\}_{i=1}^M,$$

where p_i is a paragraph, γ_i and τ_i its associated GRI and TCFD respective labels, $d(\gamma_i)$ and $d(\tau_i)$ their label descriptions, llm_i the Yes/No LLM validation output, j_sim_i the joint similarity score, c_i the final climate label, and M the number of samples after LLM validation (whereas $llm_i = Yes$).

Paragraph Extraction, Segmentation, and Pre-processing. Sustainability reports are extracted from PDF documents using layout-aware parsing with PyMuPDF. Headers, footers, and fragmented blocks are removed, with text blocks containing at least 20 words retained, as candidate paragraphs.

Weak GRI Annotation Initialization via Indexed and Non-Indexed Label Generation from the GRI Index Table. Most sustainability reports include a GRI content index mapping disclosure codes to page numbers. Although self-reported, this index provides valuable weak supervision.

For each paragraph p on page π , we define:

- **Indexed set:** GRI codes explicitly linked to π in the index.
- **Non-Indexed set:** all remaining GRI codes not mentioned in the index for π , but potentially relevant according to the semantic content.

This dual-set strategy mitigates incomplete coverage and potential strategic under-reporting.

Climate-Focused GRI-TCFD Pairing via Official Standard Mapping. To ensure conceptual consistency, we adopt the official GRI-TCFD alignment published by the GRI organization. This mapping (\mathcal{M}) restricts candidate labels to climate-relevant combinations.

We focus on seven core GRI codes (GRI 2, 201, 302, 303, 305, 306, 101), yielding 15 valid GRI-TCFD pairs. These pairs define the candidate annotation space. For each paragraph p and each candidate GRI code γ , we generate triples (p, γ, τ) where $\tau \in \mathcal{M}(\gamma)$, such as: GRI 305:Emissions \rightarrow TCFD:Governance.

Embedding-based Similarity Scoring. To rank candidate label pairs (γ, τ) , we compute semantic similarity between paragraphs and standard descriptions using the MPNet encoder (Song et al., 2020). Given the paragraph p and the texts connected to the GRI disclosure requirement and TCFD definition for γ and τ respectively, we encode them into fixed-dimensional embedding vectors. Let e_p denote the embedding of paragraph p . For each GRI code γ , let $R_\gamma = \{r_1, \dots, r_{|R_\gamma|}\}$ represent the set of official disclosure texts associated with γ . Similarly, for each TCFD category τ , let $D_\tau = \{d_1, \dots, d_{|D_\tau|}\}$ denote the set of textual definitions and guidance statements describing τ . For each triple (p, γ, τ) , we compute a joint similarity score given by:

$$j_sim_p = \max_{r \in R_\gamma, d \in D_\tau} \cos(e_p, e_r) \cdot \cos(e_p, e_d).$$

The multiplicative formulation enforces joint relevance to both reporting frameworks.

Final Label Disambiguation via Indexed and Non-Indexed Similarity Comparison. For each paragraph, we compare the highest-ranked indexed and non-indexed candidate triples. The candidate pair (γ, τ) with the higher similarity score j_sim_p is selected as the final label for each p , regardless of whether it comes from the indexed or non-indexed set. This conservative strategy prioritizes precision in the annotations.

LLM-based Validation. While semantic similarity models are powerful for linking text to structured concepts, they can sometimes overestimate relevance, especially for vague, generic, or multi-topic paragraphs. To further improve annotation quality, embedding-based similarity is followed by LLM-assisted validation using GPT-OSS 20B (Agarwal et al., 2025). Each retained triple is evaluated through a structured prompt containing the paragraph and official definitions. In line with (Mousavian Anaraki et al., 2025a), the model outputs a binary relevance decision, approximating expert judgment and substantially reducing false positives. By filtering only on “Yes” LLM outputs and using Θ_1 as the minimum allowed joint similarity threshold, we assign final climate labels to paragraphs as climate-related. By filtering only on “GRI 2 (General)” with a maximum allowed threshold Θ_2 , non-climate paragraphs can also be recognized. The resulting climate-related paragraphs constitute a reliable input for the later greenwashing risk analysis and evaluation described in the following subsection.

3.2. Greenwashing Risk Proxy Construction

Building on the climate-related annotations obtained in Section 3.1, we construct a composite greenwashing risk proxy by classifying commitment, specificity, and sentiment attributes using few-shot, in-context prompting. We emphasize that the proposed indicator represents a proxy for potential greenwashing risk based on textual characteristics, not a direct measure of deceptive corporate behavior.

To quantify the extent to which corporate climate-related disclosures may exhibit characteristics of greenwashing, we construct a composite greenwashing risk proxy based on linguistic and semantic attributes identified in prior literature (Bingler et al., 2024; Vinella et al., 2024). Bingler et al. (2024) introduces ClimateBERT-based downstream tasks to analyze firm-level climate communication. In particular, their methodology relies on fine-tuned language models to classify paragraphs according to climate relevance, sentiment, corporate commitments and actions, and linguistic specificity. These tasks are subsequently aggregated to form the

Cheap Talk Index, which measures the proportion of non-specific climate-related commitments in corporate disclosures.

Building on this framework, we adopt the same core dimensions, *commitment*, *specificity*, and *sentiment*, as fundamental components for assessing the credibility of climate-related statements. Vinella et al. (2024) suggests that greenwashing is commonly associated with three main linguistic patterns: (1) the absence of explicit climate-related commitments, (2) the use of non-specific or vague language, and (3) the use of overly positive or optimistic sentiment without corresponding substantive actions. Following this literature, we operationalize greenwashing risk by combining indicators that capture these characteristics at the paragraph level. Figure 2 illustrates the second part of our methodology for greenwashing risk proxy construction after climate relevance identification from the previous part (3.1). The proposed methodology consists of three main stages: (i) attribute-based paragraph classification via prompting, (ii) automatic dataset extension, and (iii) construction of a composite greenwashing risk proxy GWR for downstream classification and benchmark augmentation.

Attribute-Based Classification via Prompting.

For each climate-related paragraph p_i , we automatically infer three linguistic and semantic attributes: commitment Co_i , specificity Sp_i , and sentiment Se_i . These attributes correspond to key dimensions of climate communication defined in prior work (Bingler et al., 2024; Vinella et al., 2024). We design three task-specific prompts, each including two representative examples (2-shot), sampled from the corresponding ClimateBERT benchmarks, namely `climate commitments actions` (Bingler et al., 2023a), `climate specificity` (Bingler et al., 2023d), and `climate sentiment` (Bingler et al., 2023c). The prompts are constructed based on the formal task definitions provided in the benchmark documentation and are used to classify each paragraph as follows (Bingler et al., 2024):

- **Commitment:** “A paragraph gets labeled as *commitment-yes* if it reports that the company undertook activities in this regard, if it reports that it will likely do so, or if the company sets targets in this paragraph.”
- **Specificity:** “A paragraph gets labeled as *specific* if a paragraph contains detailed performance information, details of actions, or tangible and verifiable targets.”
- **Sentiment:** “A paragraph is labeled as an *opportunity* if it mainly discusses business opportunities or the positive impacts of mitigating or adapting to climate change. It is considered

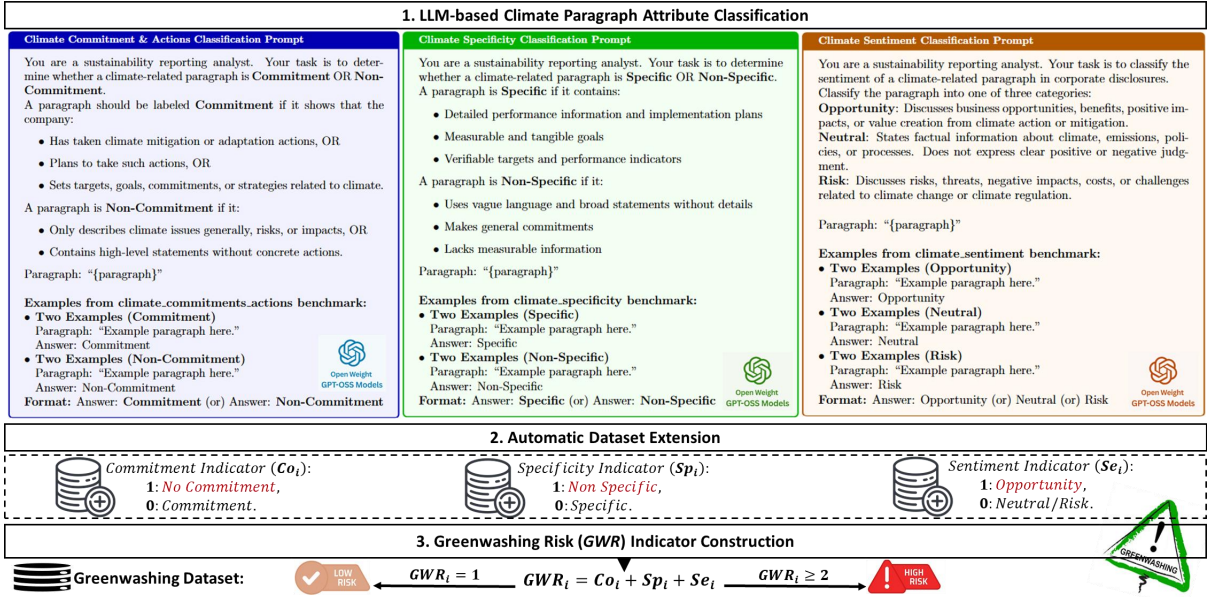


Figure 2: Overview of the greenwashing risk proxy construction pipeline based on prompt-based attribute classification, automatic dataset extension, and composite risk aggregation.

neutral if it is about facts without putting them into a positive or negative perspective. Lastly, it is labeled a risk if it mainly talks about business risks or the negative impacts of climate change.”

- **Specificity Indicator (S_{p_i}):** 1 if the paragraph is classified as non-specific, 0 otherwise.
- **Sentiment Indicator (S_{e_i}):** 1 if the paragraph exhibits optimistic framing (opportunity), 0 otherwise.

Prompt responses are used to assign categorical labels for each dimension. This process yields automatic paragraph-level annotations that are consistent with existing benchmark formulations while enabling scalable extension to large, unlabeled corpora.

Automatic Dataset Extension. Using the prompting framework described above, we extend our automatically labeled climate dataset with commitment, specificity, and sentiment annotations. This procedure enables the creation of a large-scale paragraph-level dataset without requiring manual labeling.

Greenwashing Risk Proxy Construction. Building on prior work (Vinella et al., 2024), we operationalize greenwashing risk as a composite indicator reflecting three characteristic patterns commonly associated with misleading sustainability disclosures: absence of substantive commitments, vague language, and overly optimistic framing.

For each paragraph p_i , we define three binary indicators:

- **Commitment Indicator (C_{o_i}):** 1 if the paragraph is classified as non-commitment, 0 otherwise.

The Greenwashing Risk (GWR) proxy is computed as the unweighted sum:

$$GWR_i = C_{o_i} + S_{p_i} + S_{e_i}.$$

The score ranges from 0 to 3 and reflects the number of greenwashing-related linguistic signals present in a paragraph. We discretize the score into ordinal risk categories:

$$\text{Class}_i = \begin{cases} \text{Low Risk,} & \text{if } GWR_i = 1, \\ \text{High Risk,} & \text{if } GWR_i \geq 2. \end{cases}$$

Paragraphs with $GWR_i = 0$ do not exhibit any of the targeted greenwashing patterns. Although these paragraphs are climate-related, they are excluded from the risk classification task because our indicator is a textual proxy, not a definitive measure of compliance or deception. Labeling these paragraphs as fully compliant would require additional domain-specific supervision, such as expert audits or regulatory verification, which is beyond the scope of the current framework. Focusing on paragraphs with nonzero risk scores ensures the model learns from instances where textual patterns indicate potential greenwashing, avoiding overinterpretation of compliant disclosures.

As a result of this procedure, we obtain a paragraph-level dataset consisting of climate-related disclosures annotated with ordinal greenwashing risk proxy labels (Low Risk and High Risk). This automatically constructed dataset serves as the basis for benchmark augmentation and is subsequently used for downstream evaluation of greenwashing risk classification in subsection 4.2.

4. Experimental Evaluation

Using the datasets constructed in Section 3, we evaluate the effectiveness of our annotation pipeline through indirect benchmark augmentation, i.e., by expanding manually annotated benchmark datasets with automatically generated labels and measuring downstream performance gains. This setup allows us to assess whether the proposed annotations provide useful supervisory signals beyond simply increasing the training data size.

Our evaluation focuses on climate-related and greenwashing classification tasks using ClimateBERT, a domain-adapted transformer model initialized from DistilRoBERTa and further pretrained on a large corpus of climate-related research abstracts, corporate disclosures, and news articles. ClimateBERT has demonstrated strong performance across multiple climate-related classification benchmarks (Webersinke et al., 2022). Specifically, we assess whether incorporating automatically annotated data improves downstream classification performance on established climate-related benchmarks. We consider the following benchmark settings from the ClimateBERT benchmark suite:

- **Climate Detection:** binary classification of paragraph-level `climate_relevance` dataset (Bingler et al., 2023b).
- **Greenwashing Risk Classification:** paragraph-level indicators derived from the `climate_commitments_actions` (Bingler et al., 2023a), `climate_specificity` (Bingler et al., 2023d), and `climate_sentiment` (Bingler et al., 2023c) datasets, which we transform into a unified greenwashing risk proxy (*GWR*) following the aggregation scheme described in Section 3.2.

These benchmark datasets were manually annotated by domain experts and analyzed in prior work (Bingler et al., 2024). Our greenwashing risk formulation follows Vinella et al. (2024).

We investigate the following research questions:

1. **RQ1.** Does augmenting benchmark datasets with automatically constructed annotations improve downstream performance, thereby demonstrating the informational value of the proposed pipeline?

2. **RQ2.** Does joint multi-task learning provide additional gains over separate single-task models?

Experimental Setup. We applied our pipeline to 30 sustainability reports spanning 10 industrial sectors, totaling 3,663 pages. After preprocessing, we obtained 19,133 paragraphs, of which 8,533 were associated with GRI climate-focused index entries. For climate labeling, the joint similarity threshold was set to $\Theta_1 = 0.35$, while $\Theta_2 = 0.01$ was used to identify non-climate paragraphs via the GRI 2 (General) filtering strategy described in Section 3.1. Each benchmark was expanded by approximately 50% of its original size. We fine-tuned ClimateBERT¹ on the augmented datasets and evaluated performance using Macro-averaged F1 metrics.

4.1. Climate Detection

For binary climate relevance classification, we use the `climate_relevance` (Bingler et al., 2023b) dataset. The dataset consists of 1,300 training instances (1000 climate-related and 300 non-climate-related samples) and 400 test instances (320 climate-related and 80 non-climate-related samples). All samples are written in English and were collected from corporate disclosures. We augment the training set by adding 650 samples from our dataset, resulting in a combined training set of 1,950 samples (1648 climate-related and 302 non-climate-related samples). As shown in Table 1, aug-

Training Data	Macro-averaged F1 (%)
Original (1300 samples)	93.70
Combined (1950 samples)	94.90

Table 1: Climate detection performance on the test set (400 samples).

menting the training data improves macro-averaged F1 by 1.2 points, improving overall macro-averaged F1, indicating better balance across classes.

4.2. Greenwashing Risk Classification

Following the methodology described in Section 3.2, we construct paragraph-level greenwashing risk proxy labels by combining the outputs of three ClimateBERT downstream benchmarks: `climate_commitments_actions` (Bingler et al., 2023a), `climate_specificity` (Bin-

¹Training was performed using an effective batch size of 32 and learning rate of 2×10^{-5} . The model was trained for 5 epochs using the AdamW optimizer and a linear learning rate scheduler.

gler et al., 2023d), and `climate sentiment` (Bingler et al., 2023c). We apply the same process used for our main dataset to these benchmarks to derive a unified greenwashing risk proxy. Thus, benchmark and auto-labeled samples share a consistent labeling schema grounded in established expert annotations. Paragraphs that are classified as climate-related but for which all three indicators are zero (commitment, specificity, and optimistic sentiment indicators) are not assigned a risk label and are excluded from this analysis, as these cases require additional domain-specific resources for reliable assessment. As a result, the original dataset consists of 890 training instances (516 high-risk and 374 low-risk samples) and 280 test instances (178 high-risk and 102 low-risk samples). We evaluate the impact of dataset augmentation by comparing ClimateBERT trained on the original benchmark data with a model trained on the combined training dataset (1,500 samples: 927 high-risk and 573 low-risk), which includes our automatically labeled instances. Table 2 reports the results. As shown, augmenting the training data improves macro-averaged F1 by 2.4 points, indicating that the automatically constructed annotations provide informative supervisory signals for greenwashing risk modeling.

Training Data	Macro-averaged F1 (%)
Original (890 samples)	68.10
Combined (1500 samples)	70.50

Table 2: Single-task greenwashing risk classification performance on the test set (280 samples).

4.3. Multitask ClimateBERT Fine-Tuning

We perform multitask fine-tuning of ClimateBERT to jointly model climate detection and greenwashing risk classification (capturing climate commitments, specificity, and sentiment), leveraging shared patterns to improve generalization. To accommodate the multitask framework, we introduced an extended label scheme for greenwashing risk (*GWR*) classification. Since our training data is automatically generated and may contain noise, we adopt a `prediction-based consistency constraint`: when the model predicts a paragraph as `NC`, it is encouraged to assign the `Not Consider` label in the *GWR* task. This promotes cross-task consistency while retaining flexibility and reducing spurious predictions.

The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{climate} + \mathcal{L}_{GWR} + \lambda \mathcal{L}_{consistency},$$

where $\lambda = 0.5$ controls the trade-off between task

accuracy and cross-task consistency. This consistency mechanism is designed to improve robustness and mitigate cross-task inconsistencies, particularly in the presence of noisy automatically generated labels. Table 3 summarizes the macro-averaged F1 scores across both tasks.

Training Data	Climate F1 (%)	GWR F1 (%)
Original (1190 samples)	93.89	76.16
Combined (1802 samples)	93.95	78.05

Table 3: Multitask fine-tuning results on the test set (360 samples), reporting macro-averaged F1 scores for greenwashing risk and climate relevance classification.

4.4. Results: Analysis and Discussion

From the results summarized above, we derive empirical evidence addressing our two research questions.

RQ1. Across all experimental settings, dataset augmentation with automatically labeled paragraphs leads to consistent performance gains, demonstrating the quality of these data. For climate detection, augmenting the training data increases the macro-averaged F1 score from 93.7% to 94.9% (Table 1). For greenwashing risk classification, augmentation improves macro-averaged F1 from 68.1% to 70.5% in the single-task setting (Table 2) and from 76.16% to 78.05% in the multitask setting (Table 3).

RQ2. Multitask fine-tuning further improves greenwashing risk detection by jointly modeling climate relevance and risk indicators. Compared to the single-task GWR model (Table 2), multitask training increases macro-averaged F1 for greenwashing risk from 68.1% to 76.16% on the original dataset and from 70.5% to 78.05% on the combined dataset (Tables 2 and 3), while climate detection remains stable at 94% F1, demonstrating that joint optimization does not compromise the primary classification task. The multitask setting additionally includes non-climate paragraphs, which are assigned the `Not Consider` label for the greenwashing risk task; this accounts for the difference in dataset size relative to the single-task setting.

To account for the influence of additional non-climate samples labeled `Not Consider`, performance was also evaluated excluding this class, yielding consistent gains from 68.1% \rightarrow 68.9% (original) and 70.5% \rightarrow 71.79% (combined), confirming improvements are not solely driven by label distribution. This restricted evaluation provides a fairer comparison with the single-task setting, since it focuses only on comparable climate-

related instances. Class-wise analysis shows multi-task learning primarily enhances high-risk detection (76.6% → 79.7% original; 81.6% → 81.87% combined). In contrast, low-risk performance exhibits mixed trends, slightly decreasing on the original dataset (59.5% → 58.1%) but improving on the combined dataset (59.4% → 61.7%), indicating that joint modeling better captures subtle linguistic patterns associated with potential greenwashing, particularly when supported by automatically annotated data. To better understand model behavior, we conducted a brief qualitative analysis of representative predictions made by the multi-task ClimateBERT model. In particular, we examined cases where the model predicted a paragraph as climate-related and high-risk for greenwashing (NON-COMMITMENT: Active (1), NON-SPECIFIC: Active (1), SENTIMENT: Risk (0)). One illustrative test-set example reads:

“Ecological factors and environmental regulations for access to raw material deposits also create a degree of uncertainty. In some regions of the world, for example, in West Africa south of the Sahara, raw materials for cement production are so scarce that cement or clinker needs to be imported by sea. Rising transportation costs and capacity constraints in the port facilities can lead to an increase in product costs. Overall, we rate this as a low risk.”

This example illustrates a borderline case in which the paragraph contains environmental and regulatory language that is climate-relevant, yet does not express a concrete climate commitment or measurable mitigation action. The multitask model assigns a high-risk label because the statement remains relatively generic and non-specific, which is consistent with our proxy definition of potential greenwashing risk. More broadly, such cases highlight the difficulty of distinguishing between informative contextual discussion and disclosures that may remain vague or weakly substantiated from a climate accountability perspective.

5. Conclusion

This paper proposes a scalable, paragraph-level framework for aligning corporate sustainability disclosures with GRI indicators and TCFD pillars, integrating regulatory semantics with modern NLP architectures. By focusing on paragraph-level units, the framework addresses a critical granularity gap in current reporting practices, where climate-related information is dispersed across lengthy, unstructured documents and is difficult to retrieve and interpret systematically. Our results indicate that weakly

supervised dataset augmentation improves classification robustness compared to models trained exclusively on manually annotated data. The combination of embedding-based similarity scoring with LLM-based validation enables efficient expansion of high-confidence training instances while controlling annotation noise. Fine-tuning ClimateBERT in multi-task settings further enhances generalization, indicating that climate relevance and linguistic greenwashing risk provide complementary supervisory signals. Nevertheless, given the relatively limited size of the evaluation sets, the observed improvements should be interpreted with appropriate caution. While the gains are consistent across all experimental settings, future work should validate these findings on larger manually verified benchmark datasets and through more extensive statistical significance analysis.

The introduction of a paragraph-level greenwashing proxy grounded in commitment intensity, specificity, and sentiment contributes a structured linguistic dimension to disclosure analysis. Unlike document-level ESG scoring approaches (e.g., (Wood et al., 2025)), our framework captures within-document heterogeneity, enabling fine-grained identification of potentially symbolic versus substantive climate communication. A key strength of this approach lies in its alignment with regulatory requirements. While prior studies apply LLMs to ESG or climate-related text classification, they typically treat sustainability as a thematic category. In contrast, our approach leverages the official GRI-TCFD mapping as structured supervision, thereby reflecting the architecture of climate-related financial supervision. This structured alignment enhances interpretability and comparability, and increases the potential applicability of the framework to financial risk assessment. Furthermore, by reducing reliance on exhaustive manual annotation, the proposed pipeline lowers the cost barrier for large-scale climate disclosure monitoring. This is particularly relevant for banks, regulators, and institutional investors who must process vast volumes of unstructured textual data under evolving reporting standards (Wood et al., 2025). Future research may extend this framework by incorporating additional sustainability standards and developing longitudinal models to capture the temporal dynamics. Additional directions include improving the robustness of greenwashing risk estimation through richer expert-validated supervision, broader cross-sector evaluation, and more systematic qualitative error analysis. Overall, this study contributes to the intersection of climate finance and computational linguistics by providing a replicable, standard-aligned, and economically scalable methodology for AI-assisted climate disclosure analysis.

6. Bibliographical References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. [gpt-oss-120b & gpt-oss-20b model card](#).
- Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025. Large language models for sustainability reporting: A systematic review and research agenda. *Sustainable Futures*, 10:101494.
- Julia Anna Bingler, Mathias Kraus, Markus Leipold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.
- Julia Anna Bingler, Mathias Kraus, Markus Leipold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.
- M. Bronzini, C. Nicolini, B. Lepri, A. Passerini, and J. Staiano. 2024. [Glitter or gold? deriving structured insights from sustainability reports via large language models](#). *EPJ Data Sci.*, 13(1):41.
- Eduardo C Garrido-Merchán, Cristina González-Barthe, and María Coronado-Vaca. 2026. Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related issues in corporates' financial and non-financial reports. *Neural Computing and Applications*, 38(1):12.
- Anna Glazkova and Olga Zakharova. 2025. [From data to grassroots initiatives: Leveraging transformer-based models for detecting green practices in social media](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 1–9, Tallinn, Estonia. University of Tartu Library.
- Charlott Jakob, Vera Schmitt, Salar Mohtaj, and Sebastian Möller. 2024. Classifying sustainability reports using companies self-assessments. In *Future of Information and Communication Conference*, pages 547–557. Springer.
- Agnieszka Janik and Adam Ryszko. 2025. Greenwashing in sustainability reporting: A systematic literature review of strategic typologies and content-analysis-based measurement approaches. *Sustainability*, 18(1):17.
- Hyewon Kang and Jinho Kim. 2022. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Appl. Sci.*, 12(11):5614.
- A Kazakov, S Denisova, I Barsola, E Kalugina, I Molchanova, I Egorov, A Kosterina, E Tereshchenko, L Shutikhina, I Doroshchenko, et al. 2023. Esgify: Automated classification of environmental, social, and corporate governance risks. In *Doklady Mathematics*, volume 108, pages S529–S540. Springer.
- Yao Li and Michael Rockinger. 2024. Unfolding the transitions in sustainability reporting. *Sustainability*, 16(2):809.
- W. Moodaley and A. Telukdarie. 2023a. [A conceptual framework for subdomain specific pre-training of large language models for green claim detection](#). *Eur. J. Sustain. Dev.*, 12(4):319.
- W. Moodaley and A. Telukdarie. 2023b. [Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review](#). *Sustainability*, 15(2):1481.
- Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025a. [Automatic GRI-SDG annotation and LLM-based filtering for sustainability reports](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 775–784, Cagliari, Italy. CEUR Workshop Proceedings.
- Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025b. [Unsupervised sustainability report labeling based on the integration of the GRI and SDG standards](#). In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 151–162, Vienna, Austria. Association for Computational Linguistics.
- Hui Qian Ngee, Asha Ganesh, Muhammad Aizat Noor Azmi, Tiong Yew Tang, Muaadh Mukred, Fathey Mohammed, and Adi Affandi Bin Ahmad. 2024. Environmental, social and governance (esg) scores automation in global reporting initiative (gri) with natural language processing. In *2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS)*, pages 1–7. IEEE.
- Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. [An NLP approach for the analysis of global reporting initiative indexes from corporate sustainability reports](#). In *Proceedings of the First Computing Social Responsibility Workshop*

- within the 13th Language Resources and Evaluation Conference, pages 1–8, Marseille, France. European Language Resources Association.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.*, 33:16857–16867.
- Avalon Vinella, Margaret Capetz, Rebecca Patichis, Christina Chance, Reshmi Ghosh, and Kai-Wei Chang. 2024. [Leveraging language models to detect greenwashing](#).
- Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. [Climatebert: A pretrained language model for climate-related text](#).
- Katherine Wood, Chaehyun Pyun, and Hieu Pham. 2025. Beyond green labels: assessing mutual funds’ esg commitments through large language models. *Finance Research Letters*, 74:106713.
- Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#).
- Faturahman Yudanto, Yunita Sari, and Maeve Zahwa Adriana Crown Zaki. 2024. [Climate-NLI: A model for natural language inference and zero-shot classification on climate-related text](#). In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 600–608, Tokyo, Japan. Tokyo University of Foreign Studies.
- Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. Esgreveal: An llm-based approach for extracting structured data from esg reports. *Journal of Cleaner Production*, 489:144572.
- Bingler et al. 2023c. [ClimateBERT Climate Sentiment Dataset](#). ClimateBERT Project. HuggingFace Dataset Repository. Paragraph-level climate sentiment classification dataset. Licensed under CC BY-NC-SA 4.0.
- Bingler et al. 2023d. [ClimateBERT Climate Specificity Dataset](#). ClimateBERT Project. HuggingFace Dataset Repository. Paragraph-level climate specificity classification dataset. Licensed under CC BY-NC-SA 4.0.

7. Language Resource References

- Bingler et al. 2023a. [ClimateBERT Climate Commitments Actions Dataset](#). ClimateBERT Project. HuggingFace Dataset Repository. Dataset of climate-related commitments and actions annotated from corporate disclosures. Licensed under CC BY-NC-SA 4.0.
- Bingler et al. 2023b. [ClimateBERT Climate Detection Dataset](#). ClimateBERT Project. HuggingFace Dataset Repository. Expert-annotated dataset for detecting climate-related paragraphs in corporate disclosures. Licensed under CC BY-NC-SA 4.0.

Towards Empowering Consumers through Sentence-level Readability Scoring in German ESG Reports

Benjamin Josef Schüssler[◇], Jakob Prange[♣]

[◇]University of Augsburg, Germany

[♣]German Center for Addiction Research in Childhood and Adolescence (DZSKJ),
University Medical Center Hamburg-Eppendorf, Germany
benjamin.schuessler@uni-a.de, j.prange@uke.de

Abstract

With the ever-growing urgency of sustainability in the economy and society, and the massive stream of information that comes with it, consumers need reliable access to that information. To address this need, companies began publishing so called Environmental, Social, and Governance (ESG) reports, both voluntarily and forced by law. To serve the public, these reports must be addressed not only to financial experts but also to non-expert audiences. But are they written clearly enough? In this work, we extend an existing sentence-level dataset of German ESG reports with crowdsourced readability annotations. We find that, in general, native speakers perceive sentences in ESG reports as easy to read, but also that readability is subjective. We apply various readability scoring methods and evaluate them regarding their prediction error and correlation with human rankings. Our analysis shows that, while LLM prompting has potential for distinguishing clear from hard-to-read sentences, a small finetuned transformer predicts human readability with the lowest error. Averaging predictions of multiple models can slightly improve the performance at the cost of slower inference.¹

Keywords: sentence-level readability, German ESG reports, crowdsourcing

1. Introduction

In order to make transparent how corporate economic goals align with, contribute to, or violate sustainability goals, policymakers demand written reporting on environmental, social, and governance topics, in short, ESG reports.² Next to *greenwashing* (the intentional or negligent misrepresentation of one's sustainability strategy to sound more positive and marketable than it really is, *de Freitas Netto et al., 2020*), another challenge is ensuring the reports' accessibility to their diverse audiences. This is even more important for layperson consumers than for other stakeholder groups such as economic auditors or financial analysts. The latter know exactly what they are looking for and, in case of unclear language, can consult with legal or public relations experts. This is usually not the case for consumers, who may be on their own and may read exploratorily, to gather information from scratch. Quoting EU Directive 2024/825, also known as the Empowering Consumers Directive,³ "[i]n order to contribute to the proper functioning

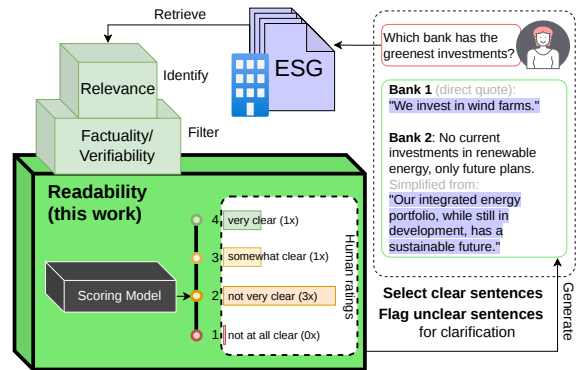


Figure 1: Readability as a foundation of consumer empowerment from ESG reports.

of the internal market, based on a high level of consumer protection and environmental protection, and to make progress in the green transition, it is essential that consumers can make informed purchasing decisions and thus contribute to more sustainable consumption patterns. That implies that traders have a responsibility to provide **clear, relevant and reliable information.**"

In this work, we focus on clarity as a fundamental requirement for consumer accessibility, and evaluate automatic readability scorers against the judgments of layperson readers (figure 1).

Automatic readability assessment (ARA) is the task of estimating how easy a text is to read and understand. Often, this is measured at the document level using broad sentence and word length

¹Code and dataset extension available at: github.com/schuesslerbenjamin/Sentence-level-Readability-Scoring-in-German-ESG-Reports. Trained models available at:

huggingface.co/schuesslerbenjamin/Sentence-level-Readability-Scoring-in-German-ESG-Reports

²EU Directive 2022/2464, also known as the Corporate Sustainability Reporting Directive (CSRD): eur-lex.europa.eu/eli/dir/2022/2464/oj/eng

³eur-lex.europa.eu/eli/dir/2024/825/oj/eng

statistics to assess generic readability. For example, the Flesch Reading Ease test (Flesch, 1948) is intended to rate English educational books on a scale ranging from school grades to professional scientist difficulty.

Here, on the other hand, we are interested in measuring readability in German ESG reports and in more fine-grained grammatical patterns than length. As our target group, we envision, for example, a young adult deciding on a sustainable bank to open their first account or investment plan with, or a family choosing an electricity provider for their home. Already overwhelmed with the multitude of companies to choose from and only able to skim very short excerpts of reports from each company, they rely on a retrieval-augmented generation (RAG; cf. Kleinle et al., 2024) or recommender system (cf. Hillebrand et al., 2023). We consider a previously unaddressed requirement for such a system, namely to maximize the clarity of the presented content: systems should (a) prefer easily readable sentences for direct extraction and (b) if a sentence is highly relevant but difficult to read, it should be simplified. Rather than trying to replicate coarse document-level scores like Flesch, we thus propose to model readability at the sentence level.

We aim to answer two questions: **RQ1:** How readable are German sustainability reports? And **RQ2:** How to model sentence-level readability? Concretely, we contribute:

- an in-depth data analysis of German ESG reports through a crowd-sourcing annotation study, finding largely clearly written sentences but also subjective variation;
- a comparison of different model types, including generative Large Language Models (LLMs), regressions, and a custom feature-based classifier, finding lower prediction error in small finetuned models and higher ranking correlation, albeit on a shifted scoring scale, in one of the LLMs;
- an ablation of syntactic features, highlighting their relative importance in sentence-level readability prediction;
- and a discussion of sentence-level readability in the context of other factors of consumer empowerment through ESG reporting.

2. Related Work

To find similar research, we systematically queried the typical research databases (see Appendix A).

Readability of German texts. While most research on readability has focused on English texts

(Collins-Thompson, 2014), some approaches have also been adapted to the German language. Amstad (1978), for example, adapts the Flesch Reading Ease formula by Flesch (1948) by changing the factor for the word length to consider that German words tend to be longer. More recent research includes creating more sophisticated readability formulae (e.g. “Hohenheimer Komplexitätsindex für Politikersprache” (HKPS, German for Hohenheim Complexity Index for Political Language) by Kercher (2013)), improving the readability for people with learning difficulties (e.g., Jablotschkin et al. (2024)), or analyzing how difficult language learners perceive the readability of texts (e.g., Weiss and Meurers (2022)).

Furthermore, the GermEval 2022 shared task on text complexity assessment of German texts by Mohtaj et al. (2022) is based on sentences from articles in the areas of society, science, and history of the German Wikipedia. It motivated a wide range of approaches, the best of which was an ensemble of GBERT and GPT-2 submitted by Blaneck et al. (2022) and achieved a 0.195 MSE (0.442 RMSE) on a 7-point rating scale.

Our study, instead, focuses on the readability as perceived by *native speakers* who are laypersons in the *ESG domain*.

Readability of ESG reports. Smeuninx et al. (2020) compare the performance of traditional readability formulae with a few modern NLP methods when predicting the readability of English ESG reports. They find that the former lack in performance, especially when the syntax varies. In general, Smeuninx et al. (2020) identify that ESG reports can be difficult to read, in some instances even more complex than financial reporting.

Among other linguistic aspects, Huang et al. (2024) analyze the readability of Chinese ESG reports and their impact on the ESG scores over time. Bonn and Gaida-Albers (2024) investigate how report readability, among other parameters, correlates with the overall sustainability of German companies (“ESG-Score” assigned by auditors), but they do not *predict* readability and the reports they analyze are written in English.

Methodologically very similar to our work are Vajjala and Meurers (2012), who compare syntactic features against “traditional features” like word length and sentence length and achieve 0.023 MSE (0.15 RMSE) on a 5-point rating scale. But they, again, work with English texts in the educational domain rather than German ESG reports.

3. Data

For our experiments, we use the dataset⁴ from the SustainEval GermEval shared task on understanding sustainability reports (Prange et al., 2025). It consists of short excerpts sampled from the German Sustainability Code (Deutscher Nachhaltigkeitskodex),⁵ where companies can voluntarily publish ESG reports and receive feedback and resources to prepare for legally required and audited CSRD reporting. Specifically, each datapoint consists of four consecutive sentences in German, of which the last is the target sentence receiving annotation and the preceding ones are provided for context. Statistics are given at the top of table 1.

Readability Annotation. We extend the “verifiability” annotations used for the SustainEval shared task with layperson readability judgments via crowd-sourcing. Training and evaluation crowd annotators were recruited via Prolific and paid above German minimum wage. The actual annotation was carried out via SoSciSurvey on GDPR-compliant servers in Germany. Development crowd annotation was carried out on a different platform, also according to German minimum wage and GDPR standards. The change in annotator pools likely led to the difference in agreement and score distributions. In all cases, the only information disclosed by annotators was that they speak German as their primary language. Annotators were identified only by anonymous IDs, which enabled us to exclude annotators from future annotation rounds if they were too fast or always assigned the same category. While the three context sentences were shown, annotators were asked specifically to rate their understanding of only the target sentence on a forced-choice Likert scale (*How well do you understand the sentence?* 1: not at all, 2: rather not, 3: somewhat, 4: very clearly).

Agreement. Most sentences were rated by 5 annotators (some by 4 and very few by 6), and most of the time (72.3–87.7%), a majority of at least 3 annotators assign the same rating (middle part of table 1). Due to the lack of annotator identities in the data, we were not able to compute chance-corrected agreement metrics such as Cohen’s κ or Krippendorff’s α . To gain a more comprehensive measure of agreement, we introduce *Mode Agreement*, which cleanly handles anonymity and varying numbers of annotations per sentence. For each sentence, we count how many annotators agree on the most common rating (the mode), and divide by the number of annotations that sentence

	Train	Dev	Eval
# Sentences	960	267	407
Ø Words / sentence	16.92	17.50	17.32
Ø Syllables / word	2.28	2.32	2.28
Inter-Annotator Agreement			
≥ 3 agree	86.8%	72.3%	87.7%
Mode agreement	70.3%	60.4%	70.1%
Readability Annotations [1.0; 4.0]			
Avg. mean	3.515	3.200	3.526
Avg. standard deviation	0.505	0.691	0.501
Avg. majority vote	3.695	3.431	3.709
# Actual majority votes			
1.0	5	0	0
1.5	1	0	0
2.0	21	11	7
2.5	19	14	4
3.0	167	90	81
3.5	76	38	35
4.0	671	114	280

Table 1: Dataset statistics.

received (see equation 1). If all annotators agree, the Mode Agreement is one. If no annotators agree, the Mode Agreement is zero.

$$\text{Mode Agreement} = \begin{cases} \frac{\text{Mode's Frequency}}{\# \text{ Annotations}}, & \text{if Mode's Frequency} \geq 2 \\ 0, & \text{else.} \end{cases} \quad (1)$$

Weighing the majority agreement in this way, while still in an acceptable range of 60.4–70.3%, paints a somewhat less optimistic picture than simply counting how often a majority exists (see table 1). Based on this, we decided to account for outlier noise in crowd-sourcing by aggregating the annotations using the majority vote instead of the mean over all votes. Only in the case of ties, we take the mean of the tied votes.

Score distribution. The bottom part of table 1 shows that the annotations are skewed towards very easily readable texts. This is most substantial in the training and evaluation splits, where more than two thirds of all instances were assigned the highest readability score of 4. This can have several reasons: Firstly, the annotators are self-declared native laypeople. We expect that most laypeople generally perceive texts in their native language as at least somewhat readable. Secondly, crowd-workers might fear that if they rate a text as not understandable, they might not be allowed to answer the other questions and get paid less. To ensure models learn to predict scores across the full scale during training, we randomly oversample the underrepresented rating classes in the training split until they match the most common

⁴github.com/SustainEval/sustaineval2025_data/

⁵deutscher-nachhaltigkeitskodex.de

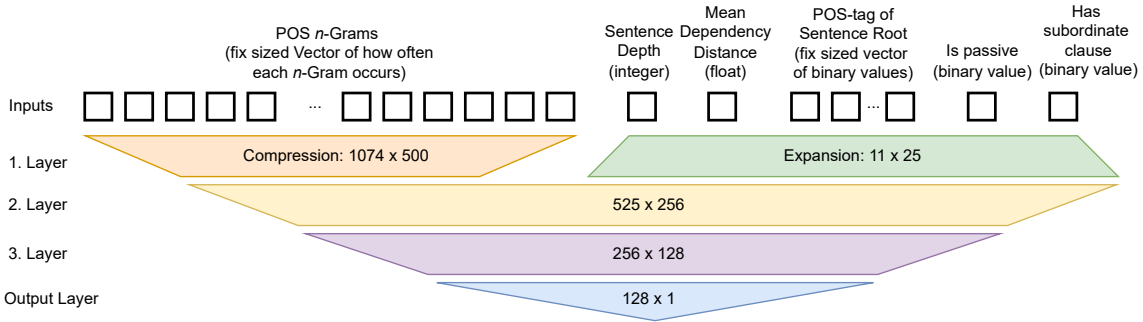


Figure 2: Structure of our syntax-based ARA model.

class. We do not manipulate the distribution in the development and evaluation splits. Scores are normalized to $[0.0; 1.0]$ for model training, inference, and evaluation.

4. Whitebox Readability Models

In line with our goal to predict sentence-wise readability in ESG reports in a transparent and interpretable manner, we explore a simple but effective whitebox model, where input features can be controlled, parameter sizes are small, and runtimes are fast (section 4.1). We hypothesize that to capture actual human judgments in a specialized domain, established formulae with linear coefficients preset to a fixed educational setting are not sufficient and test this with a baseline model (section 4.2). On the other hand, state-of-the-art language models (section 5) may be more powerful than we need and their large blackbox architectures restrict the linguistic insight they may provide.

Given the use case outlined in the introduction, where a system needs to judge how presentable individual sentences are to a layperson user, we focus on the lower-bound setup of providing only the target sentence to the models, without context. Within-sentence grammatical patterns are complementary to across-sentence semantic and pragmatic aspects of readability like verifiability, coherence, and cohesion. These aspects likely also have a large impact on readability and are modeled to various extents by pretrained (L)LMs. By examining quantitative and qualitative differences in readability scores assigned by the various models, we can approximate which facets of readability can be determined from syntax alone and which stem from other linguistic properties.

4.1. Syntactic Features

Inspired by the work of Liu et al. (2025) and Smeuninx et al. (2020), we design a feed-forward neural network on top of syntactic features extracted from the input sentence (figure 2). Most notably, to

limit the impact of the Part-of-Speech (POS)-tag n -Grams, their vector is compressed in the first layer before being concatenated with the other features. For more implementation details, see appendix B.

Our model uses the following features:

Part-of-Speech-tag n -grams. The first syntactical feature analyzes the grammatical structure of a sentence, as it can have an impact on the syntactical complexity and, thus, on the readability of a sentence (Razon and Barnden, 2015). This is based on the idea that if a sentence structure is observed more often, a reader is more likely to understand it easily (Kauchak et al., 2017). Using a sliding window, we represent a shallow view of a sentence’s syntactic structure as count features of POS-tag bigrams and trigrams. We filter for punctuation symbols. Before training, we generate all bigrams and trigrams that appear in the training set and expand them into individual features, representing how often each bigram and trigram appears in the sentence. From the training data, we extract 158 unique bigrams and 916 unique trigrams resulting in 1,074 total n -gram feature dimensions.

Depth of the dependency tree. As proposed by Yngve (1960), we calculate the depth of the dependency tree to estimate the hierarchical complexity of a sentence. In a dependency tree, every word except for the root has exactly one head that it refers to. The deeper the dependency tree of a sentence, the higher its syntactic complexity, and thus—we hypothesize—the harder it is for the reader to understand the relations of the words within the sentence.

Mean dependency distance. We also calculate the mean dependency distance as proposed by Liu (2008). The dependency distance is thereby defined as the number of words between a word and its head. Using the mean dependency distance instead of summing up all distances within a

sentence prevents longer sentences from getting disproportionately higher scores (Liu, 2008).

Part-of-Speech-tag of the root. There is exactly one word in every sentence that has no head in the dependency tree, the root. It can have a major impact on a sentence’s readability (Dell’Orletta et al., 2011). We extract all root POS-tags within the training data and expand them into binary variables. We find that verbs or auxiliary verbs are usually the root of sentences in our dataset. Thereby, our approach allows the model to find relations between all possible POS-tags of sentence roots and the readability of a sentence.

Passive voice. Sentences written in passive voice can also be harder to read. Thus, Smeuninx et al. (2020) analyze the readability of whole documents and calculate the proportion of sentences that are in passive voice. Since we are only working with single sentences, we create a binary variable indicating whether the text is in passive voice. We consider a sentence to be passive voice if it contains a participle that has a form of “werden” (the equivalent of passive *to be*) as its head, or if it includes a passivized subject.

Subordination. Finally, sentences consisting of multiple clauses can be more complicated than sentences with fewer clauses. Smeuninx et al. (2020) calculate the average number of subclause-introducing elements per sentence to represent the degree of subordination in a document. We adopt their idea to our sentence-level ARA task and create a binary variable that indicates whether there is at least one subordinate conjunction in the sentence.

4.2. Baselines

To test whether our selected linguistic features are more informative in our setting than established work on readability suggests, we compare with two baselines. Both baselines are trained on our German ESG-report data to account for domain effects.

Sentence Length. We train a simple linear regression model using only the number of words per sentence. This approach was used by Crossley et al. (2007) as a proxy for syntactic complexity.

Readability Formulae. Representing the traditional research on readability, we train an XGBoost model (Chen and Guestrin, 2016) over scores calculated using established readability formulae.⁶

⁶We also experimented with aggregating the scores using Linear Regression, Ridge Regression, Lasso Re-

We select the following formulae due to their relevance and applicability to German sentences: the *Flesch-Reading-Ease* test introduced by Flesch (1948) for English texts and adapted to German texts by Amstad (1978); the *Hohenheim Complexity Index for Political Language* (HKPS, Kercher, 2013); the proportion of polysyllabic words, based on the idea of the *Gunning Fox* (Gunning, 1952) and *SMOG indices* (McLaughlin, 1969); the *Vienna Educational Text Formula* (Bamberger and Vanacek, 1984); and the Swedish readability index LIX (Björnsson, 1968). See appendix C for details on the formulae and their implementation.

5. Blackbox Readability Models

As reference, we also compare the syntactic features model with two types of modern language models: a finetuned classifier on top of a pretrained transformer encoder and instruction-tuned generative LLMs. This is to set a practical upper bound in terms of predictive power. If a whitebox model reaches or surpasses the blackbox models’ prediction accuracy, the whitebox model should clearly be preferred. Otherwise, a tradeoff between accuracy, speed, and interpretability needs to be found.

5.1. XLM-RoBERTa Encoder-Classifier

For the first language model, we use a transformer encoder model and task-specifically finetune it to the ARA task. This approach follows Tseng et al. (2019) and can simultaneously consider several linguistic layers of a text, including semantic and syntactic aspects, making it more powerful in principle than the syntax-based model. Since we define ARA as a regression task, we train the model’s final layer as a regression head (see appendix D for more details).

This allows the model to effectively predict readability scores on a scale from zero to one. We compare several BERT-like encoder models on the development set, and select the multilingual XLM-RoBERTa-base⁷ and XLM-RoBERTa-large⁸ models (Conneau et al., 2020; Liu et al., 2019) based on their performance.

5.2. Generative LLMs

For the second language model, we test instruction-tuned LLMs on the ARA task, focusing on instruction-tuned models pretrained on datasets that include German texts.

gression, and Elastic Net, but XGBoost led to the best results overall.

⁷huggingface.co/FacebookAI/xlm-roberta-base

⁸huggingface.co/FacebookAI/xlm-roberta-large

Type	Model	MSE (\downarrow)	MAE (\downarrow)	Kendall τ (\uparrow)	\emptyset Time per Sentence (\downarrow)	# Params
□ Whitebox	Sentence length baseline	0.1859	0.4017	-0.2290	0.0003s	1
	Readability formulae baseline	0.0394	0.1588	<u>0.0863</u>	0.0018s	5
	Syntactic features (ours)	<u>0.0389</u>	<u>0.1502</u>	0.0534	0.0261s	~0.7M
■ XLM-RoBERTa	base	<u>0.0295</u>	0.1114	<u>0.2461</u>	<u>0.0035s</u>	~278M
	large	0.1325	0.3373	-0.2198	NA	~550M
■ LLMs	Qwen 3 4B Instruct 2507	<u>0.1119</u>	<u>0.2469</u>	<u>0.2822</u>	<u>0.0503s</u>	~4,000M
	Gemma 3 4B it	0.2396	0.4402	0.0448	0.3110s	~4,000M
	Llama 3 8B instruct	0.2347	0.4230	-0.1906	0.7100s	~8,000M
Combinations	Syntax + XLM-base	0.0264	0.1141	0.1676	<u>0.0296s</u>	~279M
	Syntax + Qwen	0.0485	0.1777	0.2304	0.0764s	~4,001M
	XLM-base + Qwen	0.0358	0.1432	0.2857	0.0538s	~4,278M
	Syntax + XLM-base + Qwen	0.0292	0.1355	0.2627	0.0799s	~4,279M

Table 2: Results of the experiments. \downarrow indicates that a lower value is better and \uparrow indicates that a larger value is better. The best value per metric is bold and the best per model type is underlined. For details on the experimental setup see appendix F.

We compare the Llama 3 8B instruct model⁹ by Dubey et al. (2024) with the Qwen 3 4B Instruct 2507 model¹⁰ by Yang et al. (2025) and the Gemma 3 4B it model¹¹ by Kamath et al. (2025). We analyze these LLMs, as they are highly relevant in the current research and have been extensively researched.

We prompt the models with similar instructions as the human annotators. While the dataset itself is in German, we prompt the models in English as previous research found that several LLMs are biased to internally pivot towards English due to imbalanced training data (Wendler et al., 2024). We instruct the models to classify the readability of a sentence into four classes, each coded with a number from one through four. Similar to the human annotations, these numbers are then scaled down to the same range from zero to one. We apply one-shot prompting (see appendix E), because early experiments on the development data split showed that zero-shot prompts lead to worse performance. For the shots, we randomly sample sentences and their readability score from the training data.

6. Experimental Results

All performance metrics can be found in table 2, while table 3 contains average predicted scores. Details on the experimental setup can be found in appendix F.

Metrics. We use Mean Squared Error (MSE) as our main metric for analysis and as the loss func-

⁹huggingface.co/meta-Llama/Llama-3.1-8B-Instruct

¹⁰huggingface.co/Qwen/Qwen3-4B-Instruct-2507

¹¹huggingface.co/google/gemma-3-4b-it

	Train	Dev	Eval
Sentence length baseline	2.523	2.535	2.531
Readability formulae baseline	3.527	3.551	3.528
Syntactic features	3.608	3.555	3.596
XLM-RoBERTa base	3.790	3.933	3.920
Qwen 3	3.077	3.075	3.037
Human Annotation	3.695	3.431	3.709

Table 3: Average model predictions.

tion during training. We also report the Mean Absolute Error (MAE), as it is more robust to outliers than the MSE, and more interpretable because it is true-to-scale. For both error metrics, a lower score is considered better, where 0 is the best possible error and 1 is the worst possible.

Additionally, we report a rank order correlation score to analyze whether a model can correctly identify which sentences are easier to read than others. The sentences are sorted by the predicted scores, and then this order is compared to the sorted list of gold-standard annotations (Collins-Thompson, 2014). Rank correlation measures the extent to which the predicted order aligns with human annotations across entire datasets. A model that can distinguish between easy and hard-to-read sentences but has a systematic bias to too low or high scores, provides more value than a model that makes incorrect predictions in both directions. MSE does not capture this difference, which is why a rank order correlation score is needed. We use the Kendall τ coefficient, variant b, introduced by Kendall (1945) as it is often used in the research and accounts for ties. To calculate the scores, we

use the Python implementation by the library `scipy` by Jones et al. (2001)¹². A Kendall τ score of +1 indicates perfect correlation or correct ranking, relative to the ground truth. A score of 0 indicates no correlation, and -1 shows that the rankings are inverted.

6.1. Individual Models

Sentence Length. This simple baseline yields mediocre error metrics and the Kendall τ score is worse than random, indicating that the model fails to distinguish easy from hard-to-read sentences. Further, this shows that sentence length by itself is not a good predictor of readability.

Readability Formulae. Aggregating the scores of several traditional readability formulae leads to better results than the simple sentence-length baseline, according to all metrics. This shows that the additional sentence parameters and weightings used in readability formulae allow for a better prediction of the readability than relying on the number of words alone.

Syntactic Features. With our proposed feature-based approach, we aim to predict sentence readability by having the model learn to analyze several syntactic patterns. This model is more complex than the two baselines, both in terms of its input features and degrees of freedom of its hidden layers. This added representational capacity leads to a similar error rate and ranking performance as the formulae.

To investigate the importance of individual features, we conduct an ablation study (table 4). A feature is more important to the model if its removal strongly negatively affects the performance, i.e. increases the error rate or lowers the Kendall τ score. Removing trigrams worsens the performance the most, according to all three metrics. Thus, it is the most influential feature. Passivization (by the error metrics) and the depth of the sentence (by the Kendall τ score) are, respectively, the second most important features.

Interestingly, removing the bigrams very slightly improves the performance of the model, according to the error metrics. A possible reason is that the information in the bigrams is already part of the trigrams and the model compresses the n -gram input vector to a fixed width. However, according to the Kendall τ scores, ablating any feature leads to worse performance, indicating their necessity to correctly distinguish easy from hard sentences.

¹²We also experiment with rounding the predictions to the next .5 before calculating the Kendall τ score, but as this leads to similar results, we keep the original calculation.

Ablated Feature	MSE	MAE	Kendall τ
Sentence Depth	+0.0043	+0.0282	-0.0904
Dependency Dist.	-0.0046	-0.0104	-0.0286
Sentence Root	-0.0045	-0.0321	-0.0123
Is Passive	+0.0116	+0.0329	-0.0690
Has Subordination	+0.0036	+0.0039	-0.0508
Bigrams	-0.0031	-0.0033	-0.0467
Trigrams	+0.0227	+0.0617	-0.1103
All Features	0.0369	0.1502	0.1203

Table 4: Ablation of the Syntax-based ARA model on the evaluation data split. We report the differences in metrics to the complete model (last row).

XLM-RoBERTa. We observe a strong improvement over the syntax model in all metrics. This can be argued with the transformer model’s higher degrees of freedom to fit to the task. The XLM-RoBERTa model thereby predicts with the smallest errors of all individual models. We also find, rather surprisingly, that the large model variant performs much worse than the base model variant, only slightly beating the simple sentence length baseline. This may be due to insufficient training conditions that fail to saturate the many parameters of the large model.

Generative LLMs. Prompting different LLMs, we find that Llama and Gemma fail to correctly estimate the readability, as indicated by high error metrics and very low Kendall τ scores. Qwen outperforms the other LLMs in every metric. Thus, we use Qwen as the representative LLM going forward. The highest Kendall τ score out of all the individual models indicates that the LLM is better at distinguishing easy from hard-to-read sentences than XLM-RoBERTa and the whitebox models. However, as it has not been finetuned on our specific dataset and rating scale, it is much worse than our syntax model, the readability formula model, and XLM-RoBERTa at assigning scores that are numerically close to the human ratings. Specifically, it assigns lower scores on average than most other models and humans (table 3).

Error Analysis. Consider the examples in figure 3. The top sentence is lengthy and syntactically complex. There is some variation in human judgments but consensus is clearly “not very readable”. This is reflected in all model predictions being less than the top score, though the syntax model and XLM-RoBERTa are still (too) optimistic, while Qwen matches the human vote. The bottom sentence is short and not complex, but understanding it requires access to the preceding context, which was provided to humans but not to models. This lead to a bimodal distribution in human judgments, as

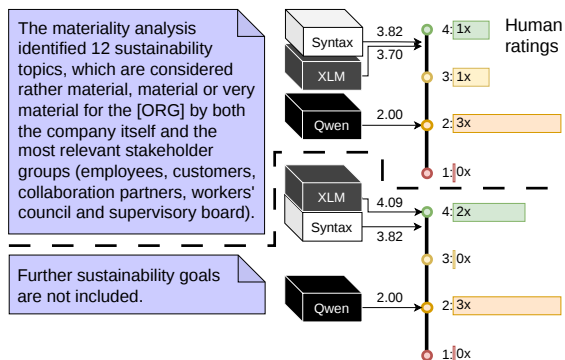


Figure 3: Two examples of different length and superficial complexity (translated from German).

some annotators likely focused on the low syntactic complexity (high readability) while others emphasized context-dependence (low understanding). Models diverge similarly, and Qwen again happens to match the majority vote which in this case is only narrowly decided. Note that Qwen and the Syntax model assign the same score to these two examples.

6.2. Model Combinations

To account for different aspects of readability influencing individual models differently, we also experiment with averaging the predictions of the three models.¹³ The combination of the syntactic model and RoBERTa has the lowest MSE (but not MAE)¹⁴ out of all experimental settings, but only by a small margin. Combining Qwen’s and RoBERTa’s predictions slightly outperforms Qwen’s individual Kendall τ score. Overall, simple mean aggregation does improve predictions slightly, but not substantially.

6.3. Trading off Errors and Speed

Some models can rate the readability of a sentence faster while making larger prediction errors than others (figure 4). Three models appear viable for this tradeoff: The readability formulae baseline is very fast while making small errors. Combining syntax and RoBERTa is slower, but makes even smaller errors. The most viable option is simply using RoBERTa, which is almost as fast as the readability formulae approach and makes almost

¹³Additionally, we tried other aggregation methods, including Linear Regression, Ridge Regression, and XGBoost. However, all models performed similarly, thus we chose mean aggregation for simplicity.

¹⁴MSE is more sensitive than MAE to individual datapoints with large errors. So the difference between the model with the lowest MSE and the model with the lowest MAE lies in deviating less from the ground truth on outliers versus getting the majority of the data closer to it.

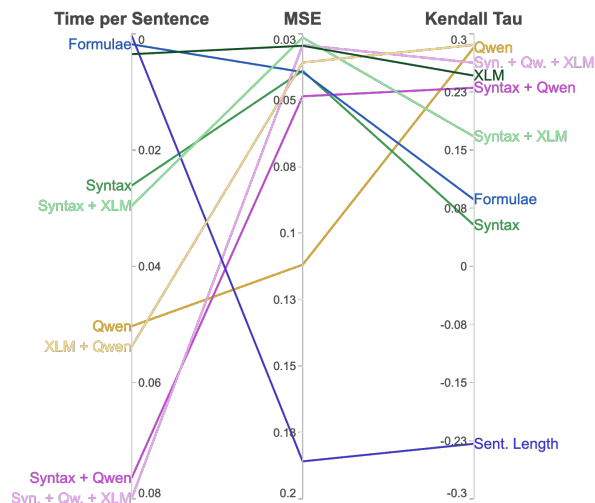


Figure 4: Time-performance trade-offs. All axes rank models top-down from best to worst.

as little errors as the combination. Any combination with an LLM has higher computational cost, slower speed, and larger score deviations.

7. Discussion

We set out to analyze how we can automatically measure the readability of German ESG reports in a way that aligns with non-expert human judgments. In doing so, we gained insight into the following two research questions:

RQ1: How readable are German ESG reports?

Our data analysis reveals that on average laypersons perceive German ESG reports as easy to read. However, some potentially crucial sentences are unclear and there is considerable variation in judgment between readers. While existing research characterizes ESG reports as generally hard to read (e.g., Pombinho et al., 2024), our sentence-level approach allows a more nuanced evaluation and, thus, enables more nuanced solutions to the problem.

In contrast to similar research that analyzes specific demographics like second language learners (e.g., Vajjala and Meurers (2012)), we have no clearly defined target audience. Our approach allows us to analyze how an average German speaker might perceive the readability of German texts. However, without a clearly defined audience, defining rules for what makes a text readable is difficult. We see this reflected in imperfect agreement among annotators. Therefore, future research could investigate personalized readability systems (Benjamin, 2012; Bailin and Grafstein, 2001) while considering that not only the grammatical structure of the target sentence, but also its

dependence on local and external context might be important.

RQ2: How to model sentence readability? We find that more complex models (more parameters) take longer to rate the readability of a sentence and tend to outperform smaller models in terms of the Kendall τ score. The LLM outperforms the other individual models in this metric, showing that it is better at delineating easy from hard-to-read sentences. However, as the LLM was not specifically tuned to the task like the other models, it performs the worst according to the MSE. XLM-RoBERTa has the best tradeoff between low MSE and fast inference.

Intuitively and according to the literature (e.g., Collins-Thompson, 2014; Vajjala and Meurers, 2012), word choice and lexical complexity play an important role as well. However, we were not able to replicate this effect in our domain and audience in pilot experiments with word frequency features.

Readability depends on the audience. A central difficulty with estimating readability is that it depends on genre and domain, as well as the audience. Traditional formulae like Flesch Reading Ease involve coefficients finetuned to the educational domain, which we account for by training a new regressor on our German ESG data. And while ESG reports are likely authored by trained writers who ensure high readability standards for expert readers, expectations may be different for the average consumer.

Empowering Consumers. As shown in figure 1, readability is a crucial building block of true consumer empowerment from ESG reporting, next to other important factors like factuality (Diggelmann et al., 2020; Florstedt et al., 2025; Luo et al., 2025) and verifiability (Prange et al., 2025). There are likely inter-correlations between these different aspects of how a company’s ESG report is written and that company’s actual sustainability strategy (Bonn and Gaida-Albers, 2024). Although the predictions of even the best models analyzed in this work are far from perfect agreement with the annotators, their scores still provide an indication for the readability as perceived by laypeople and substantiate the complexity of the task.

8. Conclusion

In this work, we applied different readability scoring methods to German ESG reports. We evaluated these methods using error and rank correlation metrics, as well as their insight into what makes a sentence hard-to-read (whitebox versus blackbox). Our results show that prompting LLMs has

the potential to distinguish clear from hard-to-read sentences. However, a small task-specifically finetuned transformer model predicts human readability with the smallest error. Averaging predictions of multiple models can slightly improve the performance at the cost of slower inference.

Feature-based models and other explainability methods, which we leave to future work, can identify individual linguistic patterns that impact readability. Thereby, future research could contribute to transparency and consumer empowerment, consumer protection and, through more sustainable consumption patterns, environmental protection and the green transition.

9. Limitations

Naturally, any model is an abstraction of reality. Thus, our models are also limited in several ways. Other hyperparameters, LMs, and more complex prompt engineering could lead to different results. Further behavioral and mechanistic explainability methods could allow more thorough investigations of the whitebox (e.g. gradient-based) and even blackbox models (e.g. discretization-based). Finally, analyzing German ESG reports on the document level could be interesting, especially regarding the coherence between adjacent sentences. Furthermore, we identify the following two major problems:

Difficulties in assessing readability. During our experiments, we find that complex context sentences can influence the perception of readability of consecutive sentences. Although the human annotators were tasked to only rate the target sentence, they were able to see the context sentences which might have impacted their ratings. However, our readability models were not able to see the context, leading to an information asymmetry. This poses a general problem to the task of sentence-level readability assessment.

Furthermore, we see a high level of subjectivity in the annotations as seen in the mediocre agreement on the readability ratings. To limit the influence of outliers, we use the majority vote to aggregate the individual annotations into a single gold truth. However, as the provided examples show, a strong disagreement can influence the majority vote drastically as well. To solve this problem, Benjamin (2012) proposes a personalized readability model trained on the user’s browser behavior. However, privacy concerns arise when tracking such personal data. We assume that for only a few users the benefits would outweigh the risks.

Class imbalance. Our data shows a strong class imbalance towards easier-to-read sentences.

Given that we only analyze the readability as perceived by native speakers, this can be argued with their fundamentally good understanding of German sentences. Furthermore, crowd-workers might fear getting rejected from the task and thus paid less if they admit to not understanding the task, or may simply overestimate themselves. This bias towards very easy sentences is in contrast to existing research that describes ESG reports as ambiguous (Bingler et al., 2024) and more complex than financial reports (Smeuninx et al., 2020). This discrepancy might have to do with our focus on sentence-level rather than document-level readability, and may either be a true effect or an artifact of how judgments were collected. We invite future research to replicate and compare different methodologies.

Future work may also address the class imbalance not only at training time but also at test time. In a simple case, for example, performance can be broken down by gold rating, evaluating instances rated as perfectly clear by all annotators separately from all instances that at least one annotator had at least some trouble understanding.

10. Acknowledgments

The authors gratefully acknowledge the HPC resources used during early experiments that were provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the BayernKI project v110ee. BayernKI funding is provided by Bavarian state authorities. The majority of the work for this paper was done at the chair of Computational Linguistics of Prof. Dr. Annemarie Friedrich at the University of Augsburg and we are very grateful for their support. Further, we thank the anonymous reviewers for their constructive feedback. We also thank Nina Prange for her input on the public communication aspect of consumer empowerment.

11. Bibliographical References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich.

Alan Bailin and Ann Grafstein. 2001. *The linguistic assumptions underlying readability formulae: a critique*. *Language & Communication*, 21(3):285–301.

Richard Bamberger and Erich Vanacek. 1984. *Lesen-Verstehen-Lernen-Schreiben*. Diesterweg.

Rebekah George Benjamin. 2012. *Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty*. *Educational Psychology Review*, 24(1):63–88.

Julia Anna Bingler, Mathias Kraus, Markus Leipold, and Nicolas Webersinke. 2024. *How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk*. *Journal of Banking & Finance*, 164:107191.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Lärarbiblioteket. Liber, Stockholm.

Patrick Gustav Blaneck, Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2022. *Automatic readability assessment of German sentences with transformer ensembles*. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 57–62, Potsdam, Germany. Association for Computational Linguistics.

Thorben Bonn and Aurin Gaida-Albers. 2024. *Does the Interaction of Informativeness, Readability, and Sentiment within Company’s Sustainability Disclosure Shape an Entity’s ESG Score? – Evidence from Germany*.

Tianqi Chen and Carlos Guestrin. 2016. *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Kevyn Collins-Thompson. 2014. *Computational assessment of text readability: A survey of current and future research*. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised Cross-lingual Representation Learning at Scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Scott A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007. *Toward a New Readability: A Mixed Model Approach*. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29(29).

Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. *Concepts and forms of greenwashing: A systematic review*. *Environmental Sciences Europe*, 32(1):19.

- Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. [READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification](#). In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Thomas Diggelmann, Jordan Boyd-Graber, Janis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A dataset for verification of real-world climate claims](#). In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.
- 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, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. [The Llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Rudolph Flesch. 1948. [A new readability yardstick](#). *Journal of Applied Psychology*, 32(3):221–233.
- Johannes Florstedt, Jonas Fahlbusch, and Moritz Sontheimer. 2025. [Detecting greenwashing in ESG reports: A comparative analysis of machine learning methods in traffic-related emissions disclosure](#). In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 25–30, Winterthur, Switzerland. Association for Computational Linguistics.
- Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.
- Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, Robin Stenzel, and Rafet Sifa. 2023. [sustain.AI: a recommender system to analyze sustainability reports](#). In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23*, page 412–416, New York, NY, USA. Association for Computing Machinery.
- Jie Huang, Derek D. Wang, and Yiyang Wang. 2024. [Textual Attributes of Corporate Sustainability Reports and ESG Ratings](#). *Sustainability*, 16(21):9270.
- Sarah Jablotschkin, Elke Teich, and Heike Zinsmeister. 2024. [DE-lite - a new corpus of easy German: Compilation, exploration, analysis](#). In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117, St. Julian's, Malta. Association for Computational Linguistics.
- Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001. [SciPy: Open source scientific tools for Python](#).
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivièrè, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce

- Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Pappas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Juyeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle K. Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Faret, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry (Dima) Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- David Kauchak, Gondy Leroy, and Alan Hogue. 2017. [Measuring text difficulty using parse-tree frequency](#). *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.
- Maurice G. Kendall. 1945. [The Treatment of Ties in Ranking Problems](#). *Biometrika*, 33(3):239–251.
- Jan Kercher. 2013. [Verstehen und Verständlichkeit von Politikersprache: Verbale Bedeutungsvermittlung zwischen Politikern und Bürgern](#). Springer Fachmedien, Wiesbaden.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Steffen Kleinle, Jakob Prange, and Annemarie Friedrich. 2024. [OMoS-QA: A dataset for cross-lingual extractive question answering in a German migration context](#). In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 231–248, Vienna, Austria. Association for Computational Linguistics.
- Fengkai Liu, Tan Jin, and John SY Lee. 2025. [Automatic readability assessment for sentences: neural, hybrid and large language models](#). *Language Resources and Evaluation*, 59:2265–2296.
- Haitao Liu. 2008. [Dependency Distance as a Metric of Language Comprehension Difficulty](#). *Journal of Cognitive Science*, 9(2):159–191.
- 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 pre-training approach](#). *CoRR*, abs/1907.11692.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Yunfang Luo, Xiling Cui, Qiang Liu, Qiang Zhou, and Yingxuan Zhang. 2025. [Identifying exaggeration in ESG reports using machine learning techniques](#). *Data and Information Management*, 9(2):100084.
- G. Harry McLaughlin. 1969. [SMOG Grading—A New Readability Formula](#). *Journal of Reading*, 12(8):639–646.
- Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. [Overview of the GermEval 2022 shared task on text complexity assessment of German text](#). In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9, Potsdam, Germany. Association for Computational Linguistics.
- Vinod Nair and Geoffrey E. Hinton. 2010. [Rectified linear units improve restricted boltzmann](#)

- machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, pages 807–814, Madison, WI, USA. Omnipress.
- Miguel Pombinho, Ana Fialho, and Jorge Novas. 2024. [Readability of sustainability reports: A bibliometric analysis and systematic literature review](#). *Sustainability*, 16(1).
- Jakob Prange, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel Neves, and Annemarie Friedrich. 2025. [Overview of the SustainEval 2025 shared task: Identifying the topic and verifiability of sustainability report excerpts](#). In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 229–238, Hannover, Germany. HsH Applied Academics.
- Abigail Razon and John Barnden. 2015. [A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 521–528, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.
- Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. [Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP](#). *International Journal of Business Communication*, 57(1):52–85.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: a simple way to prevent neural networks from overfitting](#). *J. Mach. Learn. Res.*, 15(1):1929–1958.
- Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. [An Innovative BERT-Based Readability Model](#). In *Innovative Technologies and Learning*, pages 301–308, Cham. Springer International Publishing.
- Sowmya Vajjala and Detmar Meurers. 2012. [On improving the accuracy of readability classification using insights from second language acquisition](#). In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.
- Zarah Weiss and Detmar Meurers. 2022. [Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?](#) In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do Llamas Work in English? On the Latent Language of Multilingual Transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Sondre Wold, Petter Mæhlum, and Oddbjørn Hove. 2024. [Estimating lexical complexity from document-level distributions](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6309–6318, Torino, Italia. ELRA and ICCL.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Victor H. Yngve. 1960. [A Model and an Hypothesis for Language Structure](#). *Proceedings of the American Philosophical Society*, 104(5):444–466.

12. Language Resource References

- Prange, Jakob and Jakob, Charlott and Göttfert, Patrick and Huber, Raphael and Wenzel Neves, Pia and Friedrich, Annemarie. 2025. [SustainEval 2025 Data](#). Available on GitHub.

A. Systematic Literature Search

To find existing research on the readability of German ESG reports, we queried several research databases: We searched in the ACL Anthology¹⁵, which focuses on NLP research. We queried the DBLP¹⁶, a German computer science bibliography, to include computer science research in general. We searched the EBSCOhost database¹⁷ and Scopus¹⁸ to include research from the area of business informatics. Finally, we searched the Web of Science¹⁹ as it includes research from various disciplines. All databases were queried using with the following search string:

```
(readability OR understandability OR ((text OR sentence) AND (complexity or difficulty))) AND ("Environment* Social* Governance" OR ESG OR "Corporate Social Responsibility" OR CSR OR "sustainability report" OR "company climate report") AND German
```

This search string thereby combines several terms that describe readability with terms for German ESG reports. We searched the title, keywords, and abstract fields on the 24th of September in 2025.

B. Implementation Details for the Syntax Model

We load the *de_dep_news_trf* POS-tagging model for the German language by the python library spacy. It allows extracting the POS-tags of each word, identifying the root of a sentence and its POS-tag, and extracting the depth of the dependency tree. The mean dependency distance is calculated by an extension to the spacy library called *textdescriptives*. If spacy detects a passivized subject or the sentence includes a participle with a form of "werden" as its head, it is considered passive. Finally, a sentence has a subordinate clause if spacy finds a subordinate conjunction.

The features are aggregated in a neural network. The first layer is split into two parts. In the first part, the *n*-grams vector is compressed to 500 neurons to reduce its impact on the model and the remaining features are expanded to 25 neurons for the second part of the first layer. Then, we

¹⁵Available at <https://aclanthology.org/>, last accessed Sep 24, 2025.

¹⁶Available at <https://dblp.uni-trier.de/>, last accessed Sep 24, 2025.

¹⁷Available at <https://research.ebsco.com/>, last accessed Sep 24, 2025.

¹⁸Available at <https://scopus.com/>, last accessed on 24.9.2025.

¹⁹Available at <https://webofscience.com/>, last accessed Sep 24, 2025.

concatenate the two parts and pass them to the second layer consisting of 256 neurons. The third layer compresses the model down to 128 neurons before the model outputs the regression value in the single output neuron. After each layer, except for the output layer, we add the ReLU activation function (Nair and Hinton, 2010) and 10% dropout (Srivastava et al., 2014). The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2019; Kingma and Ba, 2015). The following hyperparameters were identified using grid search: batch size: 20; training epochs: 40; learning rate: 0.01; early stopping patience: 15.

C. Implementation Details for the Readability Formulae Baseline

Based on their historical relevance and novelty, we decided to use the following models that are applicable to German sentences:

Flesch-Reading-Ease Test. Since its introduction by Flesch (1948), the Flesch-Reading-Ease test has often been used to rate the readability of English sentences (Kauchak et al., 2017). It calculates a readability score based on the number of words per sentence and number of syllables per word. Amstad (1978) recalculated its factors to fit the formula to German sentences:

$$\text{Flesch Reading Ease} = 180 - \left(\frac{\# \text{ Words}}{\# \text{ Sentences}} \right) - 58.5 \cdot \left(\frac{\# \text{ Syllables}}{\# \text{ Words}} \right) \quad (2)$$

Hohenheim Complexity Index. The HKPS (Kercher, 2013) is based on articles on politics from the German newspaper *BILD* and on dissertations on politics from PhD students. Shallow sentence and word features are weight against each other based on their importance in the two text groups. If a text is more similar to a *BILD* article it is easier-to-read for laypersons, whereas dissertations are harder-to-read for laypersons.

Polysyllabic Proportion. The idea that sentences containing many long words tend to be more complex has been often applied in research. This is, for example, one of the core ideas shared by the SMOG index (McLaughlin, 1969) and the Gunning Fox Index (Gunning, 1952). However, both are neither designed nor adapted to German texts. Therefore, we use the simple polysyllabic proportion as a feature for our readability formulae-based model and follow McLaughlin's definition of polysyllabic words as words with at least three syllables.

$$\text{Polysyllabic Proportion} = \frac{\# \text{ Polysyllabic Words}}{\# \text{ Words}} \quad (3)$$

Vienna Educational Text Formula. The Vienna formula was specifically designed for German scientific texts by [Bamberger and Vanacek \(1984\)](#). It considers the proportion of polysyllabic words, the length of sentences, long words, and the proportion of monosyllabic words. The authors supply three versions of the WSTF (Wiener Sachtext Formel in German). We use the first one as it is the most accurate one, according to the authors.

$$\begin{aligned} \text{WSTF} = & 0.1935 \cdot \text{MS} \\ & + 0.1672 \cdot \text{Average Words per Sentence} \\ & + 0.1297 \cdot \text{IW} - 0.0327 \cdot \text{ES} - 0.875, \end{aligned} \quad (4)$$

where MS is the percentage of polysyllabic words; IW is the percentage of words with more than six characters; and ES is the percentage of monosyllabic words. The scale represents the expected year of full-time education that is required to understand the text and ranges from 4 (easy) to 15 (very difficult).

Swedish Readability Index. Finally, the Läsbarhetsindex (LIX) was designed for the Swedish language by [Björnsson \(1968\)](#) and has already been successfully applied to other Germanic languages (e.g., [Wold et al., 2024](#)).

$$\text{LIX} = \frac{\text{Number of words}}{\text{Number of sentences}} \cdot \frac{\text{Number of long words} \cdot 100}{\text{Number of words}} \quad (5)$$

Long words are defined as words with more than six characters. [Björnsson \(1968\)](#) also provides a table that assigns readability classes for the score, but we use the raw LIX score to retain all information.

Model-specific implementation details. The Flesch Reading Ease and the first Vienna Educational Text Formula are calculated using the python library `textstat`. The formulae for the polysyllabic proportion, the LIX score, and the HKPS are implemented by the authors of this work. The XGBoost aggregation method is trained with the MSE objective and the following parameters: number of boosted trees: 100; learning rate: 0.1; maximum tree depth: 5. These parameters were identified using a simple grid search,

D. Implementation Details for the Experiments with XLM-RoBERTa

The RoBERTa checkpoints are loaded using the library `transformers`. The following hyperparameters were identified using grid search: batch size:

20; training epochs: 30; learning rate: 0.0001; weight decay: 0.001; gradient disabled for the first five layers.

E. LLM Prompting

The LLMs are loaded using the `transformers` library with disabled sampling for reproducibility. They are prompted to rate the readability of sentences as follows: First, a system prompt outlines the general task and indicates to the model that it is supposed to rate the readability of German sentences (see figure 5). Second, a user prompt outlines the structure in which the model is supposed to output. This includes describing the rating scale from 1 to 4 and telling to model to only output a single digit. Furthermore, the user prompt includes a single shot based on the training split of our dataset including the majority vote of the human annotators delineated by placeholder tokens. The user prompt ends with the sentence that is supposed to be rated. Third, the LLM’s output begins with a placeholder token for the score, followed by the score generated by the model.

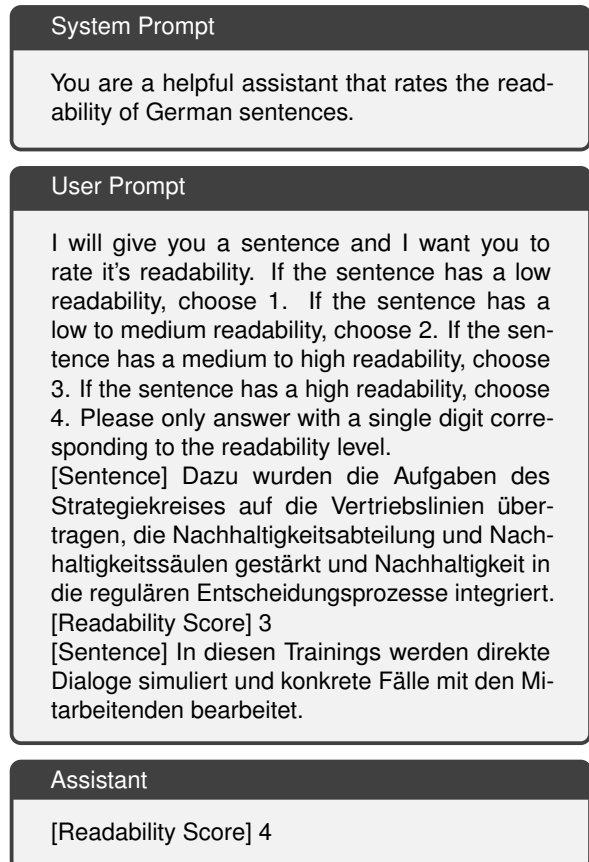


Figure 5: Prompt for the LLM-based ARA model using single-shot prompting.

F. Experimental Setup

Unless otherwise specified, we always use the default hyperparameters.

Hardware. The results to the experiments listed in this work were all created in a local workstation with an NVIDIA RTX 5080 with 16GB of VRAM, paired with an AMD 9800x3D CPU.

Software. All experiments were carried out using Python 3.12 with separate Conda environments for each model. CUDA version 12.8 was used. The experiments were run on Ubuntu 24.04 LTS within Windows Subsystem for Linux 2 on Microsoft Windows 11. We always set a seed for reproducibility.

Disambiguating Geographic Names in Biodiversity Occurrence Data: A Retrieval-Augmented Generation Approach

Yanni Jose Ella^{*◇}, Monica Ashley Laviste^{*◇}, John Michael Lastimoso[‡],
Wilfred John Santiañez[‡], Riza Batista-Navarro[†], Roselyn Gabud^{*†}

^{*} Dept. of Computer Science, College of Engineering, University of the Philippines Diliman

[‡] Gregorio T. Velasquez Phycological Herbarium and The Marine Science Institute,
College of Science, University of the Philippines Diliman

[†] Dept. of Computer Science, University of Manchester

{ycella, mrlaviste, rsgabud}@up.edu.ph, riza.batista@manchester.ac.uk,
{jmlastimoso, wjsantiañez}@msi.up.edu.ph

Abstract

The availability of georeferenced coordinates is essential for biodiversity research, as it enables species distribution modeling and supports conservation planning. However, datasets often contain ambiguous or inconsistent geographic names that reduce spatial accuracy and underscore the need for methods that resolve geographic name ambiguity. While traditional named entity linking strategies are well established, they remain limited in low-resource domains, e.g., in biodiversity contexts, due to the scarcity of annotated training data and high lexical ambiguity of local geographic names. This study proposes a Retrieval-Augmented Generation (RAG) framework to automatically disambiguate Philippine seaweed-related geographic names in databases and literature. This approach utilizes a custom knowledge base of gazetteers to support large language models (LLMs) in the task of geospatial disambiguation. With a disambiguation accuracy of 87.8% within a 5 km distance error threshold, our evaluation shows that the RAG-enabled pipeline significantly outperforms standard LLM baselines (Accuracy@5km = 0%), demonstrating the need for external knowledge to resolve geospatial ambiguity.

Keywords: retrieval-augmented generation (RAG), entity linking, large language models (LLMs), chain-of-thought (CoT), biodiversity, seaweeds

1. Introduction

Biodiversity occurrence data consists of records that capture the presence of a particular species at a defined location and point in time. The availability of geographic coordinates in these records enables species distribution mapping and habitat modeling, providing a foundational basis for biodiversity research and environmental policymaking (Chapman and Wieczorek, 2020). Despite its importance, a substantial proportion of biodiversity occurrence databases, e.g. GBIF¹, and ALA², include textual locality descriptions but lack usable geographic coordinates, particularly among historical museum and herbarium specimens. Several of these occurrence data entries contain ambiguous, outdated, or inconsistently recorded geographic names due to legacy place names, overlapping administrative units, and local vernacular name variations. As a result, data integration across biodiversity repositories is hindered, limiting the overall utility of existing datasets for research and policymaking.

This issue is particularly significant for marine macroalgae (seaweeds), especially in the Philippines, which has more than 1,065 documented seaweed taxa, making it the most diverse in terms

of seaweed resources in the western Pacific (Lastimoso and Santiañez, 2021). Since seaweeds have historically received less scientific attention than many other marine taxa (Arceo et al., 2024), accurate georeferencing is essential to providing marine scientists with reliable information on species occurrence and distribution. This enables a clearer understanding of seaweed ecology and its environmental interactions. It is especially critical for monitoring invasive species, detecting local extinctions, and assessing habitat changes driven by both local pressures (e.g., ocean acidification) and global threats such as climate change. Furthermore, precise georeferencing supports the documentation of ecological and phenological patterns through long-term monitoring, generating essential evidence for conservation planning, resource management, and sustainable use of marine resources.

Current practice resolves these ambiguities by manually interpreting textual descriptions in the literature and consulting maps or gazetteers to infer the most plausible geographic location. For example, the phrase “*Ceramiales* specimen collected at Magsaysay, Pangasinan”, contains the location mention “Magsaysay”, a common place name found across multiple provinces and municipalities in the Philippines. Using contextual cues from literature describing seaweed collection sites in Pangasinan, one can infer that it was collected

[◇]The authors contributed equally to this work.

¹<https://www.gbif.org/>

²<https://www.ala.org.au/>

in “Magsaysay Island, Bolinao, Pangasinan” and retrieve its coordinates from a gazetteer. Although effective, this is time-consuming and has become increasingly unsustainable as biodiversity data continues to grow. This underscores the urgent need for automated approaches that accurately and consistently interpret geographic references.

Traditional automated georeferencing methods rely on gazetteer lookups, rule-based parsing, or supervised named entity recognition (NER) models, which struggle with ambiguous or evolving place names and require large, domain-specific annotated datasets (Peeters et al., 2024; Marcer et al., 2021). Creating such datasets is impractical, particularly in low-resource domains like biodiversity. Moreover, effective disambiguation often depends on external knowledge not present in the text (Overell, 2011). There is thus a need for scalable approaches that can integrate new information without relying on large amounts of labeled data.

Entity linking is a natural language processing (NLP) task that maps textual mentions to the corresponding entities in a knowledge base (KB). In this study, we formulate the georeferencing problem as an entity linking task, where each location mention is systematically mapped to its corresponding real-world geographic entity in a KB, e.g., GeoNames³. Let G denote a gazetteer KB containing a finite set of geographic entities. Each entity, $g \in G$ has associated metadata (i.e., a place name, alternative names, a feature type, administrative hierarchy, geographic coordinates). Let M denote a set of geographic mentions, where each mention $m \in M$ is a text span consisting of one or more tokens referring to a real-world geographic location. The proposed entity linking system defines a mapping function $f : M \rightarrow G$.

Recent advances in NLP, particularly large language models (LLMs) combined with retrieval-augmented generation (RAG) and in-context learning, offer a promising approach to this challenge. Recent studies such as GNEMM (Zhang et al., 2025a) show that integrating retrieval with LLMs can enhance entity linking performance without reliance on large annotated datasets. However, applying such methods to biodiversity georeferencing remains underexplored.

We propose a RAG-based approach, formulated as an entity linking task, where the system retrieves candidate entities from a gazetteer and supporting passages from literature, then uses LLM reasoning with in-context learning to disambiguate mentions and select the most plausible geographic entity. We evaluate the approach on Philippine seaweed occurrence records, where ambiguity is widespread due to shared place names across administrative levels, inconsistent spelling, and legacy toponyms.

³<https://www.geonames.org/>

2. Related Work

Various tools have been developed to semi-automate the assignment of geographic coordinates to textual location data. Tools like Bio-Geomancer (Guralnick et al., 2006), GeoLocate (Wieczorek et al., 2004) and BELS⁴ (Marcer et al., 2021) primarily depend on gazetteer lookups, string matching, and rule-based parsing, often struggling with incomplete and ambiguous place names due to the lack of context (Gritta et al., 2017).

Georeferencing has also been framed as an entity linking task composed of (1) candidate retrieval from a dictionary and (2) context-based disambiguation using neural classifiers (Kolitsas et al., 2018). Transformer-based architectures, such as ReFinED (Ayoola et al., 2022), improve contextualized representations and demonstrate strong performance on general-purpose corpora or large annotated datasets. However, they depend heavily on supervised training, with performance degrading when transferred to specialized or low-resource domains (Soliman et al., 2022).

Applying entity linking to biodiversity-related geographic names introduces domain-specific challenges. Location names are inherently ambiguous and require contextual reasoning (Overell, 2011). Moreover, biodiversity records often contain fine-grained descriptions, historical place names, and landmarks (Chapman and Wieczorek, 2020). While heuristics such as population-based ranking can improve traditional georeferencing systems for well-known locations, their effectiveness diminishes when applied to local, regional, or highly ambiguous place names. As such, supervised entity linking systems trained on general-purpose corpora may not generalize well to biodiversity datasets.

Recent advances in LLMs offer an alternative via in-context learning, where pre-trained models generalize to new tasks from instructions or a small number of examples (few-shot prompting) without parameter updates. Peeters et al. (2024) demonstrate that prompt-based LLMs achieve competitive entity linking performance without extensive fine-tuning, making them viable in domains with limited labeled data. However, standard LLMs rely solely on parametric memory, making them prone to hallucination when resolving entities that are not included in their pre-training corpus (i.e., out-of-distribution entities).

RAG addresses this by combining a pre-trained LLM with a non-parametric external knowledge source (Lewis et al., 2020). Retrieving relevant candidates before generation grounds the LLM reasoning in domain-specific knowledge, mitigating hallucination. A study by Zhang et al. (2025a) inte-

⁴https://github.com/calacademy-research/bels_dockerized

grates RAG with a geographic named entity matching framework (GNEMM) that uses LLM reasoning to rank candidates by spatial and semantic similarity, though this was evaluated only on Chinese address records, limiting its generalizability.

In this work, we combine RAG with LLM-based disambiguation to address the challenges of georeferencing biodiversity occurrence data. Unlike existing supervised entity linking systems or heuristic geoparsing tools, this approach does not require domain-specific finetuning and uses a semantically enriched gazetteer and literature excerpts as a non-parametric knowledge source to handle ambiguous place mentions in biodiversity-related text.

3. Dataset

To support the development of methods for entity linking-based georeferencing, we utilized a dataset comprised of 6,865 occurrence data entries pertaining to various seaweed taxa located primarily in the provinces of Batangas and Pangasinan, Philippines. This data was collected by a group of researchers specialising in Philippine seaweeds. Each data row includes the herbarium code, administrative hierarchy, collection date, and taxonomic classification.

There are 281 unique location names in the seaweed occurrence dataset. The ground truth coordinates were established for 248 location names through a manual annotation process conducted independently by two (2) annotators. One annotator is a domain expert, a postgraduate researcher with field experience in marine science, including firsthand knowledge of the collection sites (senior annotator). The other is an undergraduate student of Computer Science (junior annotator). The annotators assigned coordinates based solely on the location name and its administrative context without considering taxon or collection date. They interpreted unstructured locality descriptions from literature, cross-referencing Google Maps to resolve place names. In addition, 86 of the locations were already documented in the laboratory’s internal geographic database of collection sites. This was accessible to the senior annotator, who had consulted it during annotation. The annotators assigned decimal degree coordinates using the World Geodetic System (WGS84) geodetic datum. Table 1 presents the number of unique location names and their corresponding occurrence records, distinguishing between those that were double-annotated and those annotated only by the senior or junior annotators.

The dataset contains location names with highly variable spatial extents. Some refer to “barangays”, the smallest administrative units in the Philippines, while others denote islands of varying sizes, or higher-level administrative divisions such as towns.

Table 1: Number of location name annotations and the corresponding number of occurrence records.

Annotator	Nbr. of Unique Location Names	Nbr. of Occur. Records
Sr Annotator	123	2654
Jr Annotator	22	506
Both Annotators	103	3040
Not Annotated	33	665
Total	281	6865

Since the dataset contains occurrences of marine (seaweed) species, several location mentions were assigned coordinates near or along coastlines by the senior annotator, drawing on prior knowledge of the collection sites. For example, the location mention “Santiago Island, Pangasinan” may be annotated by one annotator with coordinates on the central part of the island, while another may assign coordinates on the northern part. Although the coordinate pairs differ, both fall within the geographic extent of the same locality, as seen in Figure 1.

To assess the reliability of the annotations, we measured inter-annotator agreement (IAA) on the doubly annotated set composed of 103 unique location names, treating the senior annotator’s coordinates as the reference standard. Given the variability in spatial extent, particularly in marine science research, coordinate agreement cannot always be determined through exact matching. Instead, it is more appropriately evaluated using kilometer-radius matching, with a threshold that may be larger than what is typically applied in georeferencing residential addresses. We applied a kilometer-radius matching criterion of 1, 3, and 5 km, where the annotators were considered to be in agreement if the distance between the coordinates they provided fell within the specified radius. The IAA, measured in terms of accuracy, is 50.98%, 79.41%, and 93.14% for the 1-, 3-, and 5-km radius, respectively, indicating a high level of consistency between annotators and demonstrating annotation reliability.

Because a single location name may correspond to multiple occurrence records, each unique location name was annotated only once, and the resulting coordinates were applied to all records sharing that location name. After removing entries without ground truth coordinates, i.e., entries whose location names could not be resolved to coordinates either by the laboratory’s existing database or manual annotation, 6,186 records remained. This dataset was partitioned into training, development, and test splits using an 80:10:10 ratio. The development set was used for LLM parameter tuning and retrieval configuration experiments, while the test set was strictly reserved for final evaluation.

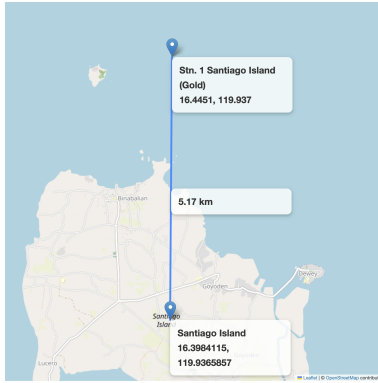


Figure 1: 5-km distance between the coordinates of the collection site “Station 1, Santiago Island” and of “Santiago Island” as seen in Google Maps.

4. Methods

In this section, we present our methods for linking a location mention to its corresponding real-world geographic entity. The input to the pipeline is a *query text*, q , formulated as a natural language representation of an occurrence record that includes the administrative hierarchy, collection date, and taxonomic classification. For example,

“*Bryopsidales Caulerpaceae (Chlorophyta, exsiccatae*; Apr 21, 1983) collected from Long Beach, Bolinao, Pangasinan, Philippines on 21 April 1983. Specimen associated with herbarium code MS113509, recorded as *Caulerpa racemosa*.”

The overall methodology has two phases. The first phase is the construction of the KB by embedding the gazetteer and seaweed literature corpus into a vector database. Figure 2 shows this process which is done only once. The second phase is the per-query pipeline, as shown on Figure 3, which comprises (1) the two-stage retrieval that takes q as input and retrieves gazetteer data and literature excerpts relevant to q , referred to as *candidates*, which are then reranked and passed to the LLM; and (2) LLM-based resolution that utilizes in-context Learning and Chain-of-Thought prompting. Here, the LLM evaluates the *candidates* against the provided context to resolve the location mention to specific gazetteer coordinates.

4.1. Gazetteer Knowledge Base

We constructed a Philippine gazetteer KB by obtaining country-specific data from GeoNames to enable *candidate* retrieval and LLM grounding. It uses decimal degrees and WGS84 as the coordinate system and geodetic datum, similar to the format of the coordinates in our dataset. It contains 96,643 Philippine geographic entities, each associated with the corresponding coordinates, ad-

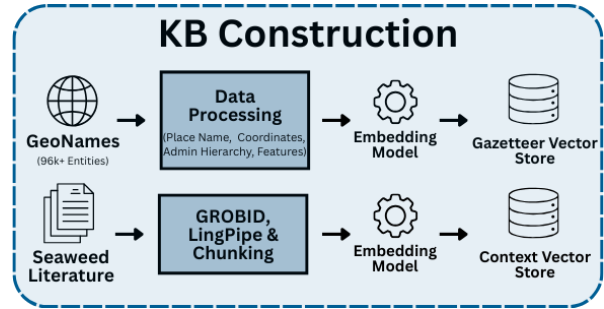


Figure 2: Knowledge Base Construction Process

ministrative classifications, and feature codes. The administrative units were standardized to reflect country-specific terminology (e.g., `ADM1` mapped to `Region`, `ADM2` to `Province`).

Each GeoNames entry was transformed into a canonical textual description capturing the place name, feature type, administrative hierarchy, and alternate names (when available). By converting structured data into unstructured text, we are able to capture the relational context between the place name and its geographic features, allowing the vector database to represent semantic meanings. This representation constitutes a single entry in the gazetteer KB and enables the retriever to match not only on string similarity but also on the semantic properties of geographic features. For example, it facilitates disambiguation of “Pasig” as either Pasig River (a hydrographic feature) or Pasig City (an administrative unit). Figure 4 shows an example textual description formed based on GeoNames attributes.

4.2. Context Knowledge Base

We constructed a context KB based on 24 scholarly articles authored by members of a research group with expertise in Philippine seaweeds, which contain ecological descriptions, sampling site details, geographic and taxon mentions (e.g. *Bryopsidales Caulerpaceae*) that may provide supporting evidence for geographic disambiguation. Text was extracted from the PDF documents using GROBID⁵, which produces structured and normalized XML representations. The extracted text was segmented into sentences and grouped into overlapping chunks using a sliding window strategy. We experimented with different chunk sizes and overlap settings to determine the configuration that best balances contextual coverage and embedding quality (see Section 5.2).

4.3. Retrieval

The system generates a dense *embedding of the query text*, $e(q)$, for the retrieval stage. Querying

⁵<https://github.com/grobidOrg/grobid>

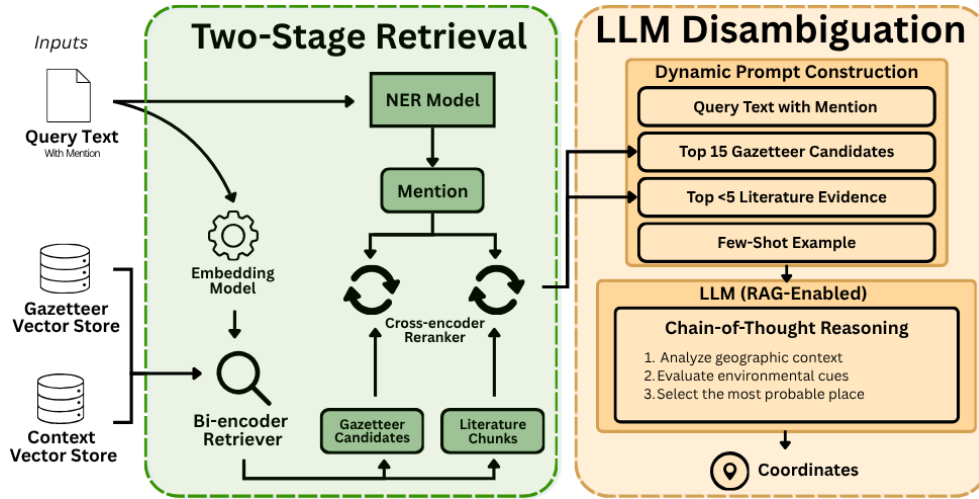


Figure 3: RAG pipeline with retrieval, few-shot prompting, and LLM generation for geographic name disambiguation.

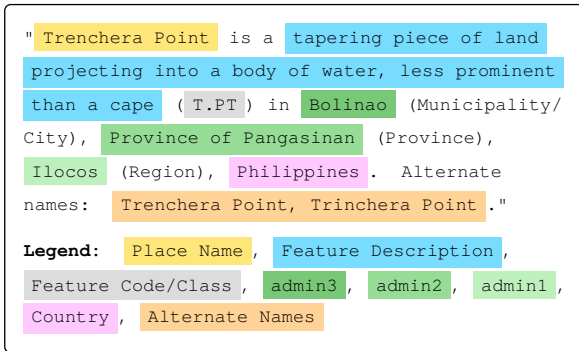


Figure 4: Example concatenated text based on a GeoNames gazetteer entry.

the gazetteer KB, the retriever computes similarity scores between $e(q)$ and each gazetteer entry embedding $e(g)$ using the cosine similarity function $s(a, b)$ and returns $C_G = \text{Top}_n\{g \in G \mid s(e(q), e(g))\}$, where Top_n returns the $n = 150$ gazetteer entities with the highest similarity scores with $e(q)$. In parallel, C_L , the set of top $k = 15$ semantically relevant chunks from the context KB, is retrieved. Both retrievals use a bi-encoder approach, in which the query and each candidate are independently encoded into dense vectors. Candidate ranking is based on cosine similarity between these vectors which enables the system to identify relevant *candidates* even if the mention does not exactly match the gazetteer entries (e.g., partial names).

To determine the most effective retrieval configuration, we evaluated five (5) dense embedding models: BGE-M3 (Chen et al., 2024), MiniLM (Wang et al., 2020), Qwen3-Embedding-0.6B (Zhang et al., 2025b), Contriever (Izacard et al., 2021) and MPNet (Song et al., 2020). These models were indexed across three (3) vector databases: ChromaDB (Xie et al., 2023), Pinecone (Xie et al., 2023) and FAISS (Johnson et al., 2021).

The bi-encoder retriever encodes queries and gazetteer entries independently, which enables nearest-neighbor search over precomputed embeddings but may not capture fine-grained interactions between the *mention* and each *candidate*. To address this, we applied a cross-encoder reranking stage using *bge-reranker-base*, which jointly encodes the *mention-candidate* pair to produce more discriminative relevance scores (Wu et al., 2020). This enables the reranker to focus on token-level differences in context (e.g., distinguishing “Patar, Bolinao, Pangasinan” from “Patar, Tayug, Pangasinan”) that the bi-encoder alone cannot resolve.

Reranking is also applied to the retrieved literature chunks, with a relevance threshold of 45% to filter out low-scoring literature chunks and to reduce the noise passed to the LLM. After reranking and filtering, the top 15 gazetteer *candidates* and top 5 literature chunk *candidates* are retained and passed to the LLM-based disambiguation stage.

4.4. LLM-based Disambiguation

For each geographic mention, a prompt is dynamically constructed containing the *query text*, q , the gazetteer *candidates*, and literature chunk *candidates*. This grounds the LLM’s reasoning on both geographic and ecological evidence.

Chain-of-Thought (CoT) Prompting. The system employs a structured CoT strategy that explicitly instructs the LLM to follow a three-step reasoning process before answer generation. The first step (Place Analysis) interprets the mention in context and is characterized solely from the *query text* q . In the next step (Candidate Evaluation), the LLM compares each retrieved *candidate* with the context, weighing administrative-region consistency, name similarity, and environmental consistency. Lastly

(Selection & Validation), the LLM returns the best-fitting *candidate* and outputs coordinates exactly from the gazetteer, preventing data hallucination.

Few-Shot Prompting. To guide the LLM’s output structure and reasoning style, the pipeline employs few-shot prompting, where the number of examples (shots) is set to 2. These examples demonstrate the resolution of the location name with its corresponding context and the expected output. This helps in anchoring the LLM’s expectations, reduces the likelihood of formatting errors and ensures that the model prioritizes administrative consistency (e.g., matching “Pangasinan” with “Ilocos Region”) over superficial name matches. Furthermore, this provides the LLM with a reasoning template for handling potential orthographic and typographical errors in the source literature and place name mentions.

5. Results and Analysis

We conducted experiments to identify an optimal configuration that balances accuracy with computational efficiency, such as runtime and model size, using evaluation metrics from georeferencing and NLP literature. We then report the results based on the manually labeled held-out test set. All experiments were conducted on an NVIDIA A100 Tensor Core GPU, and the reported results represent the average over three independent runs.

5.1. Evaluation Metrics

Due to variability in the extent of the locality, we evaluated the system using accuracy within a distance threshold. The Maximum Uncertainty framework proposed by [Thapa and Bossler \(1992\)](#), which has since been the best practice for georeferencing ([Chapman and Wieczorek, 2020](#)), explicitly accounts for uncertainty arising from the spatial extent of a locality (e.g., Santiago Island with a diameter of 7 km). This aligns with the point-radius method, which represents the best estimate point together with an uncertainty radius that bounds a plausible true location ([Wieczorek et al., 2004](#)). In addition, we used geodesic distance to measure the distance between two coordinates to account for the shape of the earth with as much precision as possible.

In this study, we report the accuracy (Acc) at multiple radii using thresholds of 1, 3 and 5 km (Acc@k km) to accommodate the varying spatial granularity of locations in our dataset, which include specific sampling sites, barangays (ADM4), municipalities, and entire islands.

To complement this, we also compute the Root Mean Squared Distance (RMSD) as a measure of

Table 2: Impact of chunking configuration (no. of sentences - overlap) on retrieval performance.

Chunk	@1↑	@3↑	@5↑	Med↓	RMSD↓
3s-1o	27.6	54.2	72.8	2.44	116.80
5s-1o	27.8	54.6	73.0	2.23	115.40
5s-2o	26.0	52.8	68.8	2.33	136.50
7s-2o	26.2	52.5	67.5	2.44	126.70
10s-3o	24.5	51.0	68.8	2.31	103.67

geographic error. This captures the overall magnitude of spatial deviation and penalizes larger errors more heavily. Lower RMSD values indicate predictions that, on average, are closer to the gold standard coordinates ([Leidner, 2007](#)). Given the geoid distance $\Delta(.,.)$ between a set of N ground-truth centroids $\vec{y} = (y_1, \dots, y_n)$ and chosen location centroids $\vec{d} = (d_1, \dots, d_n)$ retrieved from the gazetteer:

$$\text{RMSD}(\vec{d}, \vec{y}) = \sqrt{\frac{1}{N} \sum_{i=1}^N [\Delta(d_i, y_i)]^2}$$

In addition, we report the median (Med) defined as the middle value of the ordered distance errors to represent the typical precision of the system that is resilient to outliers (e.g., >100 km distance errors).

5.2. Chunking Configuration Analysis

We evaluated a total of 20 retrieval configurations across 5 embedding models: BGE-M3, Contriever, MiniLM-L6-v2, MPNET-Base-v2, and Qwen3-0.6B, while remaining within their respective token limits. We tested 3, 5, 7 and 10 sentences per chunk with overlaps of 1 to 3 sentences. However, due to token size limitations, errors occurred with all-MiniLM-L6-v2 for the 5-, 7-, and 10-sentence chunks with two or more overlaps (i.e., 5s–2o, 7s–2o, and 10s–3o). Similarly, all-mpnet-base-v2 and facebook/contriever encountered errors under the 10-sentence, 3-overlap (10s–3o) configuration.

Table 2 illustrates the retrieval performance of each unique chunking configuration on the different embedding models averaged together on a fixed LLM backbone (Gemma2:9b) and vector database (FAISS). Although the 10-sentence window yielded the lowest average error at 103.67km, it was only tested on two embedding models because of token limits. The 5-sentence window with 1-sentence overlap achieved the highest accuracy and the second lowest average error at 115.4km. Consequently, we adopted the 5s-1o configuration for the remainder of our experiments to ensure consistent performance and a broader embedding model compatibility.

5.3. Retrieval Configuration Analysis

Table 3 shows the comparative performance of embedding models and vector databases under

Table 3: Retrieval performance of embedding models and vector databases using Gemma2:9b.

DB Emb	@1↑	@3↑	@5↑	Med↓	RMSD↓
FAISS					
Bge	24.5	52.4	69.6	2.30	71.00
Contriever	25.8	24.0	79.7	1.54	49.00
MiniLM	26.9	51.5	68.7	2.30	185.00
Qwen	18.3	41.7	60.2	3.08	122.50
MPNet	21.7	43.3	61.3	3.08	138.89
Pinecone					
Bge	21.9	39.6	58.6	3.07	118.66
Contriever	24.2	52.9	69.6	2.37	136.24
MiniLM	25.2	37.8	55.0	3.49	203.33
Qwen	18.3	41.1	59.6	3.07	124.58
MPNet	21.8	30.8	49.0	5.57	180.20
Chroma					
Bge	22.9	45.9	46.3	5.81	158.48
Contriever	19.2	57.1	73.2	2.25	149.56
MiniLM	24.5	51.0	67.5	2.41	198.00
Qwen	17.4	34.4	52.4	4.56	206.70
MPNet	18.5	37.8	55.8	3.07	136.79

Table 4: Embedding model with 15 candidates retrieval comparison using FAISS vector database.

Model	@1↑	@3↑	@5↑	Med↓
Bge	38.5	71.6	84.6	1.48
Contr	39.4	95.3	96.6	1.20
MiniLM	36.0	61.0	77.3	1.95
Qwen	22.6	57.5	76.0	2.14
Mpnet	25.6	58.8	66.1	2.14

a fixed Gemma2:9B backbone and the optimal 5s-10 chunking configuration (Section 5.2).

FAISS consistently achieved the lowest RMSD across all configurations, particularly when paired with Contriever (RMSD = 49km). This represents a significant improvement over Pinecone and ChromaDB, where the same model produced substantially higher RMSD ranging from 136.24 and 149.56, respectively.

In addition, Table 4 highlights a performance gap among embedding models when their retrieval performance is evaluated on the *candidate* with the lowest distance error among the 15 retrieved results. At the 5 km threshold, Contriever achieved the highest accuracy at 96.6%, while MPNet reached only 66.1%, a disparity of more than 30%. Interestingly, BGE ranks second to Contriever in terms of performance. This can be attributed to BGE’s strong multilingual capabilities, which enable the LLM to deduce orthographic variations in Filipino, such as resolving the mention “Balahibong Manok” to the correct gazetteer entry “Balahibongmanoc” by relating semantic terms to geographic entities. Overall, Contriever paired with FAISS emerged as the superior configuration, reaching a peak Acc@5km of 96.6% with a median (Med) error of 1.20 km.

The superior performance of FAISS is mainly attributed to its exhaustive search capability. Using an IndexFlatL2 mechanism, FAISS performs

Table 5: LLM comparison using the optimal FAISS-Contriever retrieval configuration.

Model	@1↑	@3↑	@5↑	Med↓	RMSD↓
Gemma2-9B	25.8	62.5	87.8	1.54	49.30
Qwen2.5-7B	22.6	55.2	75.2	1.95	89.30
Randomblock	20.3	52.6	80.8	2.15	110.96
Llama3-8B	21.9	62.0	86.5	1.77	104.68

a brute-force search that guarantees the retrieval of the mathematically closest candidates from the 96,643-point Philippine gazetteer. Meanwhile, the Contriever’s success reveals that its unsupervised contrastive pre-training allows it to generalize effectively by focusing on raw semantic structures. This enables the system to match physical descriptions such as “coastal reefs” or “Sargassum beds” between the literature and the gazetteer metadata with high precision.

5.4. Comparison of LLMs

Using the optimal FAISS-Contriever configuration, we benchmarked different LLMs with similar parameter sizes on resolution accuracy (see Table 5). Gemma2:9b achieved the best overall performance with the highest Acc@1km (25.8%), Acc@3km (65.5%) and Acc@5km (87.8%), lowest median error (1.54km) and RMSD (49.30km). This suggests that Gemma2:9b is the most reliable in selecting the correct candidate from the retrieved set.

5.5. Ablation Study

To quantify the individual contribution of each pipeline component, we conducted an ablation study and reported their performance in Table 6.

First, we implemented a gazetteer-only baseline. This simulates the conventional approach of consulting gazetteers to resolve place names, selecting the top-ranked candidate by exact string matching of the place name, without any LLM reasoning. The configuration achieved an Acc@5 of 52.4% but had large distance errors, resulting in an RMSD of 110.40 km. The high RMSD value highlights the limitations of string matching without contextual reasoning. This is due to the system’s inability to distinguish between identical or near-identical place names that refer to geographically distinct entities.

Next, we developed an LLM-only baseline configuration using the best-performing retrieval configuration (FAISS & Contriever) and LLM (Gemma2:9b). This LLM-only configuration that relies entirely on the LLM’s parametric knowledge without any retrieval failed completely across all accuracy thresholds (0%). This confirms that parametric knowledge alone is insufficient to resolve place names in low-coverage areas and could lead to LLM hallucinations. The high median distance (67.25%)

Table 6: Ablation study with the best LLM and retrieval configuration. Gaz: Gazetteer; Lit: Literature.

Config	@1↑	@3↑	@5↑	Med↓	RMSD↓
Gaz only	23.5	40.1	52.4	1.94	110.40
LLM	0.0	0.0	0.0	67.25	70.61
+CoT	0.0	0.0	0.0	15.75	68.30
+Gaz	24.0	64.6	81.0	1.88	49.50
+Gaz+Lit	25.8	64.4	87.8	1.54	38.43

and RMSD (70.61km) indicate random or weak region-based predictions.

Using CoT prompting with the base LLM significantly reduced the median error from 67.25 km to 15.75 km, demonstrating that structured reasoning instructions help the model produce more coherent geographical outputs. However, the accuracy remains at 0% for all thresholds, suggesting that reasoning alone cannot compensate for the absence of external geographic knowledge.

Adding gazetteer retrieval to the LLM produces the largest performance gain, increasing Acc@5km from 0% to 81.0%, and reducing RMSD from 68.30km to 49.50km. This confirms that constraining the LLM’s output to retrieved candidates is the primary reason for the improvement. The gazetteer serves as a non-parametric memory that prevents hallucinations by forcing selection from valid gazetteer entities.

Finally, incorporating the literature grounding on top of the entire pipeline yielded improvements on the system’s performance. Acc@5km increased from 81.0% to 87.8%, the median error decreased from 1.88 km to 1.54 km, and RMSD dropped from 49.50 km to 38.43 km. Although the gains are smaller compared to the gazetteer component, this suggests that the literature grounding provides useful additional context in resolving the mention.

In summary, while LLMs utilize their vast general knowledge to identify and disambiguate place names with 20-50 km distance errors, they struggle with the precision required for specialized domains like country-specific seaweed research.

6. Discussion and Error Analysis

We now examine the system’s success and failure cases, as summarized in Table 7.

Success Cases. The system effectively handles spelling and naming variations that may occur in biodiversity literature. For example, the mention “Trinchera”, despite the spelling variation, is resolved to “Trenchera Point”. The dense retriever successfully identified this candidate despite the lexical mismatch. The LLM selected “Trenchera Point” based on its classification as a coastal hydrographic feature, rather than selecting the ad-

ministrative entity “Trinchera”, which occurs in two municipalities within the same province. This is explicitly stated in the retrieved literature, providing additional context. A similar case is “Balahibong Manok Island”, which was correctly resolved to “Balahibongmanoc Island”. This shows that dense embedding retrieval captures semantic similarity beyond exact string matching.

The case of “Patar” illustrates the role of the context KB in resolving intra-provincial ambiguity. Multiple gazetteer entries share the name “Patar” within the same province. In this instance, retrieved literature describing seaweed collection activities specifically in Patar, Bolinao, enabled the LLM to distinguish the correct entity from its administratively equivalent counterparts.

Failure Cases. A primary source of error and a key contributor to the RMSD of 49.30 km, is the absence of the correct entity in the gazetteer. When the correct entity is not present in the retrieved *candidates*, the system is constrained to select the closest *candidate* that is either semantically or administratively related. For example, “Maahas Vicinity” was resolved to “Maahas” within the correct province but under a different municipality. More importantly, “Tomasa” was resolved to “Doña Tomasa”, which is in an entirely different region. Despite the LLM’s own reasoning and flagging this match as “geographically impossible”, it still outputs the associated coordinates since it is designed to always select a candidate from the retrieved set. A confidence-based abstention mechanism that allows the system to return no prediction when the retrieval quality is low could address this issue.

Errors can also arise from misleading geographic signals in the retrieved literature chunks, particularly when taxon mentions greatly influence retrieval relevance. Because the context KB consists of papers with broad taxonomic coverage, some papers are retrieved and reranked higher primarily due to the taxon mention. These passages may contain places entirely unrelated to the query but highly relevant to the taxon. An example of this is “Botong” in Batangas. The retrieved literature contains a description of seaweed collection in Pangasinan, which the LLM treated as key geographic evidence. Since no gazetteer candidate located in Pangasinan was retrieved, the LLM attempted to find the geographically closest place to the region. In this case, it resolves it to “Botong” in Visayas, which is farther from Pangasinan. This suggests that the LLM lacks reliable spatial reasoning and is reliant on its parametric knowledge, leading to incorrect inference. Thus, for the literature grounding to be effective, the passages must contain information that is relevant to the query and supports the disambiguation task at hand.

Mention → Gazetteer Match	LLM Reasoning
<i>Trinchera</i> → Trenchera Point (Bolinao) SUCCESS: SPELLING VARIATION	Literature mentions seaweed collection and selected this as a coastal landform consistent with this context.
<i>Balahibong Manok Island</i> → Balahibongmanoc Island (Batangas) SUCCESS: NAME VARIATION.	Selected based on name similarity and island classification. Literature lacked geographic cues, but parametric knowledge supported the match.
<i>Patar</i> → Patar (Bolinao) SUCCESS: MULTI-MATCH	Literature confirms Bolinao, Pangasinan as the collection area.
<i>Maahas Vicinity</i> → Maahas (Los Baños) FAILURE: OUT-OF-GAZ.	Literature mentions Batangas. Selected Maahas in Laguna since it is in the same region, and the correct entity is not in the gazetteer.
<i>Tomasa</i> → Doña Tomasa (Albay) FAILURE: OUT-OF-GAZ.	Context mentions Bolinao, Pangasinan; candidate is a barangay in Albay (Bicol Region), which is geographically inconsistent.
<i>Botong</i> → Botong (Bohol) FAILURE: MISLEADING LIT.	Literature mentions Pangasinan. Selected candidate in Bohol, believing it is closest to Bolinao in distance.

Table 7: Success and failure cases of resolution.

7. Conclusions and Future Work

This paper presented a retrieval-augmented entity linking approach for georeferencing biodiversity occurrence records that contain ambiguous place names. Our framework combines a gazetteer, a context KB derived from seaweed literature, dense retrieval with reranking and LLM-based disambiguation using structured prompting. Experiments show that retrieval quality is the main driver of performance. We also note that an LLM with no retrieval support is not viable for georeferencing due to its low accuracy, but adding gazetteer and literature retrieval yields the largest performance jump. By grounding decisions in external knowledge and restricting output to retrieved candidates, the system reduces hallucination and improves geographic accuracy.

We have identified a few potential directions for future work. First is improving candidate coverage and ensuring that the retrieved context is informative and relevant. We suggest incorporating a confidence-based abstention mechanism that can give the model the option to not return a prediction when retrieval quality is low. This approach could reduce forced match errors, especially in cases where the locality is not found in the gazetteer. Experimenting with the relevance threshold to select *candidate* literature chunks from the KB can ensure relevant literature grounding. Second is the use of a domain-specific gazetteer that can help improve the coverage of biodiversity sampling sites, particularly coastlines and marine locations. This will reduce the likelihood of resolving to coordinates that are based on geopolitical boundaries instead of the coastlines or the sea. Our third recommendation is to handle uncertainties based on locality type, wherein uncertainty thresholds vary depending on whether the mention refers to an exact place or a broader region. Finally, performance evaluation can be improved by exploring additional metrics that penalize lexically similar but geographically incompatible matches. Distance-based thresholds penalize predictions only when they fall outside the

uncertainty threshold. There are instances when distance alone may not capture whether the correct entity was selected for places that are near to each other.

8. Acknowledgments

The author acknowledges the Office of the Chancellor of the University of the Philippines Dili-man, through the Office of the Vice Chancellor for Research and Development, for funding support through the Outright Research Grant (262622 ORG). This work was also partially supported by an Early Career Research Fellowship 2024-25 programme grant (“Enhancing Environmental Resilience through AI-driven Analysis of Unstructured Data”) from the International Science Partnerships Fund (ISPF) delivered by the British Council.

9. Bibliographical References

- Hazel Arceo, Joyce Velos, Ma Nuñez, and Porfirio Aliño. 2024. *The West Philippine Sea: State of the Coasts*.
- T. Ayoola, J. Fisher, and A. Pierleoni. 2022. Improving entity disambiguation by reasoning over a knowledge base. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2899–2912. Association for Computational Linguistics.
- Arthur Chapman and John Wieczorek. 2020. *Georeferencing Best Practices*. Publisher: GBIF Secretariat.
- Jianlv Chen, Shitao Xiao, Peitian Hou, Quanyue Ye, Haotian Zhang, Huaying Cao, Chao Sun, Xueying Liang, and Zhiyong Cao. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text retrieval. *arXiv preprint arXiv:2402.03216*.

- Milan Gritta, Mohammad Taher Pilehvar, Nut Lim-sopatham, and Nigel Collier. 2017. What’s missing in geographical parsing? *Lang Resour Eval*, 52(2):603–623.
- Robert P Guralnick, John Wieczorek, Reed Beaman, Robert J Hijmans, and the BioGeomancer Working Group. 2006. [Biogeomancer: Automated georeferencing to map the world’s biodiversity data](#). *PLOS Biology*, 4(11):1–2.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. [Towards unsupervised dense information retrieval with contrastive learning](#). *CoRR*, abs/2112.09118.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. [End-to-end neural entity linking](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.
- John Michael Lastimoso and Wilfred John Santiañez. 2021. [Updated checklist of the benthic marine macroalgae of the philippines](#). *Philippine Journal of Science*, 150:29–92.
- Jochen L. Leidner. 2007. [Toponym resolution in text](#). *ACM SIGIR Forum*, 41:124.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Arnald Marcer, Elspeth Haston, Quentin Groom, Arturo H. Ariño, Arthur D. Chapman, Torkild Bakken, Paul Braun, Mathias Dillen, Marcus Ernst, Agustí Escobar, David Fichtmüller, Laurence Livermore, Nicky Nicolson, Kaloust Paragamian, Deborah Paul, Lars B. Pettersson, Sarah Phillips, Jack Plummer, Heimo Rainer, Isabel Rey, Tim Robertson, Dominik Röpert, Joaquim Santos, Francesc Uribe, John Waller, and John R. Wieczorek. 2021. [Quality issues in georeferencing: From physical collections to digital data repositories for ecological research](#). *Diversity and Distributions*, 27(3):564–567.
- Simon Overell. 2011. The problem of place name ambiguity. *SIGSPATIAL Special*, 3:12–15.
- R. Peeters, A. Steiner, and C. Bizer. 2024. Entity matching using large language models. *arXiv preprint arXiv:2310.11244*.
- H. Soliman, H. Adel, M. H. Gad-Elrab, D. Milchevski, and J. Strötgen. 2022. A study on entity linking across domains: Which data is best for fine-tuning? In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 184–190.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [Mpnet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.
- Khagendra Thapa and John Bossler. 1992. Accuracy of spatial data used in geographic information systems. *Photogrammetric Engineering and Remote Sensing*, 58(6):835–841.
- 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 *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- John Wieczorek, Qinghua Guo, and Robert Hijmans. 2004. [The point-radius method for georeferencing locality descriptions and calculating associated uncertainty](#). *International Journal of Geographical Information Science*, 18:745–767.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of EMNLP 2020*, pages 6397–6407.
- Xingrui Xie, Han Liu, Wenzhe Hou, and Hongbin Huang. 2023. [A brief survey of vector databases](#). In *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, pages 364–371.
- W. Zhang, S. Chen, J. Li, and C. Xu. 2025a. Geographic named entity matching and evaluation recommendation using multi-objective tasks: A study integrating a large language model and retrieval-augmented generation. *ISPRS International Journal of Geo-Information*, 14(3):95.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#).

Appendix

The Chain-of-Thought Prompt structure as well as the sample Few-shot prompt can be seen in Table 8.

Table 8: Prompts for the Large Language Model.

Prompt Strategies	Prompt Content
Chain of Thought	ROLE: You are an expert in Philippine place name disambiguation, geography, and historical gazetteers. Use step-by-step reasoning to resolve place names to coordinates. TASK: Determine which candidate correctly identifies this place mention. Step 1: Place Analysis. Characterize the mention by its place type... Step 2: Candidate Evaluation. Multi-factor comparison... Step 3: Exact Validation. Select only from the candidates' coordinates..
Few Shot Prompting	Example A (Geographic Disambiguation): Context describes a Sargassum bed in a coastal Philippine municipality; among candidates, the Ilocos municipality best fits this description.... Example B (Geographic context over lexical match): Mention: "Panaun Island", Context explicitly states Bolinao, Pangasinan ; Candidate 1: "Panaun Island" in Southern Leyte is geographically impossible despite lexical match; Candidate 2 "Panaon" is in Pangasinan and matches context; Candidate 2 is chosen...

Sentiment and Stance in EFL Responses to AI-Generated Environmental Content

Andry Sophocleous

University of Nicosia

46 Makedonitissas Avenue, 2417, Nicosia, Cyprus
sophocleous.a@unic.ac.cy

Abstract

Recent advances in generative AI have enabled the large-scale production of environmental imagery and descriptions, yet questions remain regarding how such content represents emotion, agency, and responsibility. This study examines how human evaluators respond to AI-generated environmental representations, focusing on sentiment, stance, and argumentation as dimensions of qualitative evaluation. Data were collected from 81 multilingual secondary-school EFL learners in Cyprus, who engaged with AI-generated environmental images and accompanying AI-written descriptions through a sequence of structured tasks. Using qualitative discourse analysis informed by sentiment- and stance-oriented frameworks, the study analyses learner-produced texts to identify affective evaluations, moral positioning, and alignment with or challenge to AI-generated discourse. Findings suggest that participants consistently moved beyond surface-level description to articulate emotional engagement, assign responsibility, and critique omissions in AI-generated content, particularly regarding the representation of human-environment relations. The study contributes to research on human-centered AI evaluation by demonstrating the value of sentiment and stance analysis for assessing AI-generated environmental language and highlights the potential of educational contexts as sites for examining human interpretive responses to automated discourse.

Keywords: Ecolinguistics, Evaluation of AI-generated Environmental Discourse, Sentiment Analysis

1. Introduction

Recent advances in generative artificial intelligence have enabled the automated production of environmental imagery and written descriptions that depict animals, ecosystems, and ecological crises. Such AI-generated content is increasingly available in educational, media, and public domains, shaping how environmental issues are framed, interpreted, and evaluated. However, while these systems are often presented as neutral, their representations may implicitly privilege anthropocentric perspectives, downplay emotion, or weaken questions of agency and responsibility in human-environment relations.

From an ecolinguistic perspective, language plays a central role in constructing how the more-than-human world is perceived and valued. Discursive patterns shape emotional engagement, ethical judgement, and ultimately how animals and ecosystems are treated (Cook, 2015; Stibbe, 2012; 2015). Research in environmental communication similarly demonstrates that affective framing plays a crucial role in motivating attention, engagement, and action, with negative or emotionally charged representations often eliciting stronger responses than neutral ones (Brader, 2006; Bachleda et al., 2020; Soroka et al., 2019; Hart et al., 2024). In this sense, environmental imagery functions as a powerful affective trigger that invites sentiment, stance, and argument (Hoewe & Ahern, 2017; Wong-Parodi & Feygina, 2021).

Sentiment analysis (SA) has been widely applied in domains such as economics, healthcare, and governance to examine how people emotionally evaluate discourse topics, products, and policies (Ibrohim et al., 2023). In environmental contexts, SA has been used to analyze public perceptions of environmental conditions and climate-related policies, particularly through large-scale social media and public datasets (Du et al., 2019; Stede & Patz, 2021). However, much of this work has focused on automated or large-scale analyses, offering limited insight into how affect, stance, and moral evaluation are constructed through human interpretation. Recent work in environmental NLP has begun to address some of these limitations through the development of domain-specific language resources that extend beyond polarity-based sentiment analysis. For example, Grasso et al. (2024) introduce EcoVerse, an annotated Twitter dataset designed for eco-relevance classification, environmental impact analysis, and stance detection, highlighting the need for specialized annotation schemes to capture evaluative and ideological dimensions of ecological discourse. However, the examination of how AI-generated environmental representations are interpreted or evaluated by human readers is still an area that remains underexplored.

2. Research Gap and Study Aims

Despite growing interest in sentiment analysis and affective NLP, comparatively little attention has been paid to how humans evaluate AI-generated environmental discourse itself, particularly in terms of emotion, agency, and

responsibility. Existing studies have largely prioritized computational detection of sentiment or large-scale public opinion data (Du et al., 2019; Ibrohim et al., 2023), leaving underexplored the qualitative processes through which AI-generated representations are interpreted, contested, or reframed by human readers. Moreover, the emergence of generative AI introduces a new discourse agent whose outputs may downplay emotion or moral responsibility, potentially silencing the voices of more-than-human beings by privileging anthropocentric framings (Cook, 2015).

Educational contexts provide a productive yet underutilized site for examining such human-centered evaluation. When learners engage critically with AI-generated content, they not only respond emotionally but also articulate ethical judgements and evaluative stances that reveal how environmental meaning is negotiated. Addressing this gap, the present study adopts a qualitative discourse-analytic approach to examine how multilingual English as a Foreign Language (henceforth EFL) learners respond to AI-generated environmental imagery and descriptions. Focusing on sentiment, stance, and argumentation, the study aims to contribute to research on human-centered evaluation by highlighting affective and ethical dimensions of AI-generated environmental language that are often overlooked in computational approaches. Specifically, the present study is guided by the following three research questions:

1. How do human evaluators express sentiment when responding to AI-generated environmental imagery and descriptions?
2. What stance-taking and argumentative patterns emerge in human evaluators' written responses to AI-generated environmental representations, particularly in relation to agency and responsibility?
3. In what ways do human evaluators align with or challenge AI-generated environmental discourse with respect to emotion, moral evaluation, and the representation of human-environment relations?

3. Data and Methods

3.1 Context and Participants

The study was conducted between November 2025 and February 2026 in a secondary school in Nicosia, Cyprus. Participants were 81 multilingual adolescents (aged 15-17) from a range of L1 backgrounds, including Greek, Cypriot Greek, Bulgarian, Romanian, and Slovak, enrolled in five EFL classes. All participants demonstrated intermediate to upper-intermediate proficiency in English (CEFR B1-B2) and were accustomed to producing short written texts as part of regular classroom instruction. In connection with the scope of the study, participants functioned as

human evaluators of AI-generated environmental content. Their multilingual backgrounds provided a reflective and critical perspective on the interpretation of AI-generated imagery and text, making the setting suitable for examining human-centered evaluation of automated discourse.

During the first two weeks of the study, the preliminary phase, learners completed a set of short activities designed to introduce key concepts from ecolinguistics in an EFL context. These activities focused on word ecology and nature metaphors in English, the ecological reworking of metaphors, learners' personal relationships with nature, and cross-linguistic comparison of nature metaphors in Greek and English. This preparatory phase aimed to familiarize students with ecolinguistic perspectives and sensitize them to the role of language in shaping ecological meanings, thereby preparing them for the image-based study tasks that followed.

After completing the preliminary phase (Task 1), learners engaged in three further tasks, introduced sequentially to explore the three research questions (see Table 1). In some instances, tasks extended to two or three teaching periods (of 45 minutes each). Tasks were completed either individually or in pairs, depending on the task, allowing for both individual reflection and collaborative meaning-making in response to the image-based prompts.

TASK	FOCUS	DESCRIPTION	ANALYTICAL CONTRIBUTION
1	Ecolinguistic Awareness	Introduction to ecolinguistics through nature metaphors	Conceptual preparation (data not analyzed)
2	Seeing the Invisible	Learners described generated images	Sentiment (RQ1), initial stance
3	Voices of the Earth	Learners wrote short dialogues between human and non-human animals	Sentiment (RQ1), stance (RQ2), agency
4	Evaluation of AI Description	Learners evaluated AI-generated descriptions,	Stance & argument (RQ2), AI alignment/challenge (RQ3)

Table 1: Overview of the Four Tasks on Sentiment, Stance, and Argument

3.2 Data Materials

The materials consisted of four DALL·E-generated environmental images depicting contrasting ecological narratives: a polar bear stranded on ice, a polluted river, a thriving tropical forest, and an urban rooftop garden (see Figures 1-4 below). Figures 3 and 4 were also used as visual prompts in another study by the author. However, the present study is based on a different dataset, analytical framework, and research focus.



Figure 1: Polar Bear Stranded on Ice



Figure 2: Polluted River



Figure 3: Tropical Forest



Figure 4: Rooftop Urban Garden

These images were generated by DALL·E (OpenAI) and the following prompt: Can you generate an image of a _____? Tasks

progressed from description to emotional evaluation, stance-taking, and critical reflection. In Task 2, learners described the generated images using the prompt ‘Can you describe this image?’, which was also the prompt used with ChatGPT 5.2 to generate the descriptions which learners later evaluated in Task 4 (see Extract 6). In Task 3, learners wrote short dialogues from the perspective of humans and non-human animals in response to a whale soundtrack, encouraging empathy and voice. In Task 4, learners evaluated AI-generated descriptions of images they had not previously encountered in Task 2, identifying points of agreement, disagreement, and omission. Overall, the tasks elicited affective and evaluative language and supported creative, critical engagement with environmental issues in the EFL classroom.

3.3 Data Collection Tasks

Data were collected during regular EFL lessons across multiple sessions through classroom tasks based on visual or audio prompts. Learners worked individually or in pairs, producing written descriptions, emotional and stance-based responses, empathic dialogues, and evaluations of AI-generated image descriptions. All responses were collected in class, anonymized, and organized by task format before being analyzed as qualitative discourse data. In Tasks 2 and 4, learners worked in pairs and each pair received only one of the four images. The images were therefore not presented sequentially or simultaneously to the same participants. In Task 2, pairs described their assigned image, while in Task 4 they evaluated an AI-generated description of a different image. This design limited opportunities for direct comparison across contrasting ecological visuals and reduced the likelihood that responses were shaped by prior image exposure or responses from other groups. In Task 4, learners received the same instruction used to generate the ChatGPT 5.2 output: ‘Can you describe this image?’

3.4 Analytical Framework

Data were analyzed using qualitative discourse analysis, informed by research on evaluation, affect, and stance in discourse (Hunston & Thompson, 2000; Martin & White, 2005) and interpreted within an ecolinguistic framework focused on care, harm, agency, and the more-than-human world (Stibbe, 2021). Learners’ written responses

were analyzed as whole texts, with closer attention to salient discourse segments. The coding process combined deductive and inductive approaches: initial categories were derived from the research questions and the study's conceptual framework (e.g. sentiment, stance, responsibility, and ecolinguistic positioning), while additional patterns and subcategories emerged through close reading of the learners' responses. For RQ1, the focus was on emotional expression; for RQ2, on stance, moral evaluation, and responsibility; and for RQ3, on how learners aligned with or challenged AI-generated descriptions. Across all analyses, attention was also given to broader ecolinguistic framings, such as anthropocentric versus ecocentric positioning and care versus catastrophe narratives. The aim was to identify salient discourse patterns and interpretive tendencies rather than quantify linguistic features or assess grammatical accuracy.

4. Results

4.1 Sentiment in Learners' Responses to Environmental Imagery

Learners' responses to the environmental images revealed a range of emotional evaluations, including negative, positive, and mixed sentiment, reflecting the affective contrast embedded in the visual prompts. Negative sentiment was particularly salient in responses to the polar bear imagery (Extract 1 and Figure 1).

1) I feel anger, disgust and responsible and hopeless

2) The emotions are obviously negative

3) The emotions in the image is isolated on ice cap without any help, trying to save her life. The ice is melting while the ice time past. We feel responsible ^{disgust and} for this situation while is humans fault as we constantly destroy everything. We also feel sadness because we know that is not easy to change this situation. as a result polar bears ~~are~~ will die.

Extract 1: Describe Figure 1

Negative sentiment was expressed in responses to crisis-oriented imagery through clustered emotions, empathy for non-human suffering, and explicit moral judgement, often linked to human responsibility. By contrast, positive sentiment appeared in responses to care-oriented images (as in Figure 3), where learners emphasized calmness, balance, and emotional restoration through sensory descriptions of intact natural environments (see Extract 2).

3) This image evokes calmness because of the lush greenery, soft sunlight, and gently flowing stream. The natural setting feels peaceful and balanced, with birds moving freely through the forest, which suggests harmony and life without urgency. This emotion is helpful because it encourages relaxation, reduces stress, and creates a sense of mental clarity, as if inviting the viewer to slow down and breathe in a tranquil, restorative environment.

Extract 2: Describe Figure 3

Interestingly, numerous responses regarding the same image expressed mixed or ambivalent sentiment, combining appreciation with concern for future loss (see Extract 3).

3) This image makes me feel happy but also sad at the same time. Happy because there are still places like this on earth, untouched and beautiful, and sad at the same time because ~~probably~~ they probably will not continue existing in the future. ~~The~~ Humanity's selfishness and greed are destroying places like this ~~continuously~~ continuously. Big companies and governments also are very responsible because of their factories and ~~power~~ bad use of power.

Extract 3: Describe Figure 3

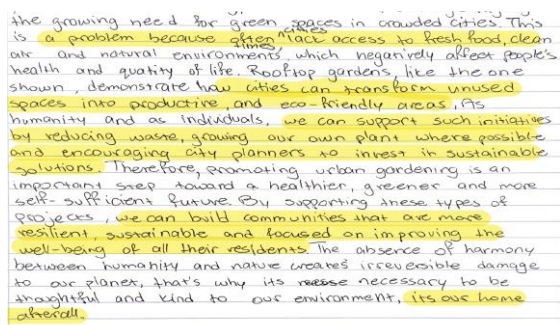
As noted in this extract, the image made the learner feel 'happy but also sad at the same time', celebrating the existence of untouched natural spaces while simultaneously expressing anxiety about their destruction due to 'humanity's selfishness and greed'. Such ambivalence illustrates how learners negotiate hope and despair within the same response, often transitioning from emotional evaluation to ethical critique.

4.2 Stance and Argumentation in Learners' Environmental Discourse

Across responses, participants demonstrated recurring stance-taking patterns that moved from emotional reaction to moral judgement and proposed action. A common argumentative sequence was observed: identification of an environmental problem, attribution of responsibility, and articulation of a normative response or call for action. Interestingly, this structure appeared across both crisis- (Extract 1) and care-oriented (Extract 4) imagery, suggesting a stable evaluative framework rather than image-specific reactions.

Stance was linguistically realized through moral evaluation (e.g. explicit judgements of harm or neglect), modality (e.g. expressions of obligation or necessity), and collective positioning (e.g. inclusive references to 'we' or 'humanity'). Responsibility was predominantly attributed to humans, with participants framing environmental degradation as the result of collective choices rather than natural processes. Even in responses

to positive imagery, participants frequently extended their evaluations to include ethical obligations to preserve or replicate such environments. This is clearly indicated in Extract 4, after students were asked to describe Figure 4.



The growing need for green spaces in crowded cities. This is a problem because often, lack access to fresh food, clean air and natural environments which negatively affect people's health and quality of life. Rooftop gardens, like the one shown, demonstrate how cities can transform unused spaces into productive and eco-friendly areas. As humanity and as individuals, we can support such initiatives by reducing waste, growing our own plants where possible and encouraging city planners to invest in sustainable solutions. Therefore, promoting urban gardening is an important step toward a healthier, greener and more self-sufficient future. By supporting these types of projects, we can build communities that are more resilient, sustainable and focused on improving the well-being of all their residents. The absence of harmony between humanity and nature creates irreversible damage to our planet, that's why it's ~~more~~ necessary to be thoughtful and kind to our environment. It's our home after all.

Extract 4: Interpreting Underlying Human-Nature Narratives

This writing shows clear stance-taking through moral evaluation, modality, and the attribution of responsibility. The student explicitly frames the issue as a problem affecting people's health and quality of life, while also presenting urban transformation as a shared ethical responsibility. The closing statement, 'it's our home after all,' further personalizes environmental responsibility and highlights human connection with the wider world. As this example shows, a problem-responsibility-action pattern appeared in responses to both crisis-oriented and care-oriented images, with learners moving from description to ethical judgement and calls for change.

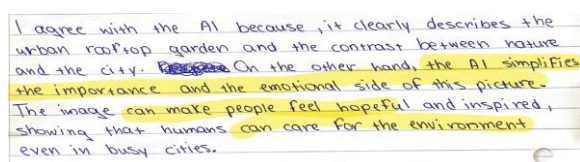
4.3 Evaluation of AI-Generated Environmental Descriptions

Learners' evaluations of AI-generated image descriptions showed both agreement with and resistance to automated environmental writing, especially in relation to emotional depth, agency, and moral responsibility. In the following example, a student responded to the following AI-generated description of the Urban Garden (Extract 5 and Extract 6):

The image shows a large rooftop garden located on top of a tall building in a modern city. Wooden garden beds filled with green plants, vegetables, and herbs are arranged in neat rows. Several people are working in the garden, watering plants and taking care of the crops. Tall glass buildings surround the rooftop, showing a contrast between nature and the urban environment. The

space looks organized, clean, and productive. The image suggests that nature and city life can exist together through planned and sustainable actions.

Extract 5: Chat GPT 5.2 Generated Text using prompt 'Can you describe this image?' (see Figure 4)



I agree with the AI because, it clearly describes the urban rooftop garden and the contrast between nature and the city. ~~Because~~ On the other hand, the AI simplifies the importance and the emotional side of this picture. The image can make people feel hopeful and inspired, showing that humans can care for the environment even in busy cities.

Extract 6: Evaluation of AI-Generated Description

In their responses to Task 4, learners generally accepted the factual content of the AI-generated texts but consistently noted the absence of emotional engagement, explicit responsibility, and attention to human – non-human relationships. They often challenged these omissions by supplementing AI-generated interpretations with affective language, explicit blame, or ethical commentary. In doing so, they repositioned animals and ecosystems as sentient and morally significant entities rather than passive and secondary elements. This pattern suggests that human evaluators actively compensated for what they perceived as affective or ethical gaps in AI-generated environmental discourse. Overall, the findings suggest that human evaluators engage critically with AI-generated content, using sentiment and stance as evaluative lenses through which automated anthropocentric representations are assessed, reframed, or resisted.

5. Discussion

The findings of this study demonstrate that AI-generated environmental imagery can function as a powerful pedagogical and analytical prompt for eliciting affective, ethical, and evaluative written discourse in the EFL classroom. Across tasks, learners moved beyond surface-level description to provide emotional responses, moral evaluations, and assign responsibility in relation to ecological issues. As regards ecolinguistic concerns, learners' written discourse revealed a consistent tendency to highlight human accountability and to reposition non-human

animals and ecosystems as morally salient entities rather than secondary and passive beings. This suggests that image-based tasks can provide a productive site for observing how emerging environmental voices are articulated through learners' developing EFL skills.

This positioning relates directly to the notion of AI critical literacy, as students were encouraged to question the apparent neutrality of AI-generated environmental discourse and to reflect on what was emphasized, omitted, or depersonalized. By engaging critically with AI outputs, learners demonstrated an emerging awareness of how automated systems can shape environmental meaning, emotional engagement, and responsibility attribution, rather than simply reflecting objective reality.

It is important to note that in Task 2, both learners and ChatGPT responded to the same prompt 'Can you describe this image?' Under these comparable conditions, learners often produced more affectively and ethically loaded responses than the model, introducing emotional language, responsibility, and moral evaluation beyond neutral visual description. Yet, this should be interpreted cautiously, as different model versions may generate different kinds of output. Consider, for example, the text generated by Theophrastus 5.2, a H4rmony-instructed assistant that operates through ecosophical guidance, describing Figure 4 with the same prompt:

Theophrastus 5.2: 'Can you describe this image?'

The image depicts a vibrant rooftop urban garden set amid a dense cluster of modern skyscrapers, where several people are tending to neatly arranged raised beds filled with a rich diversity of vegetables, herbs, and flowering plants. The wooden decking and planters create a warm, human-scale environment that contrasts with the surrounding glass and steel architecture, while watering cans and gardening tools suggest ongoing care and cultivation. From an ecolinguistic perspective, this scene visually communicates an alternative "story we live by," one that repositions humans not as dominators of urban space but as participants in ecological systems, integrating food production and biodiversity into the built environment. It embodies a

narrative of reconnection, resilience, and local sustainability, aligning with the kind of life-affirming discourse advocated in ecolinguistic frameworks, where urban spaces are reimagined as sites of coexistence between human activity and ecological flourishing.

Theophrastus 5.2 produced a more interpretive and affectively loaded description than ChatGPT 5.2, which resembles students' writing as it highlights reconnection between humans and plants as participants in a larger ecological system. This suggests that AI models may vary not only in descriptive detail but also in the extent to which they introduce ethical, relational, and ideologically framed meanings.

6. Conclusion

The study demonstrates the value of educational settings as sites for human-centered evaluation of generative systems. When participants are invited to engage reflectively with AI outputs, they produce evaluative discourse that reveals interpretive strategies, ethical orientations, and expectations that may not be captured only through automated analysis. Despite the limited scope of the study and its context-specific sample, it offers insights into how human judgement can inform the assessment of AI-generated environmental language. By accentuating affective and ethical dimensions of evaluation, the findings highlight the importance of integrating qualitative human perspectives into the development and evaluation of generative language systems, particularly in domains with significant social and environmental implications. Moreover, the comparison between ChatGPT 5.2 and Theophrastus 5.2 further suggests that AI-generated environmental discourse should not be treated as homogeneous, since different models may yield different affective, relational, and evaluative framings depending on how they have been developed and fine-tuned.

7. Limitations and Ethical Considerations

This study is limited by its qualitative scope and focus on a single educational context. Hence, its findings may not be generalizable to other populations or contexts. In addition, the analysis is restricted to written responses and does not capture evaluative processes that may emerge in other modes of interaction. As regards the ethical considerations of the study, informed consent was obtained in line with institutional guidelines, and all data were anonymized prior to analysis. AI-generated content was used solely as an object of critical reflection and image generation, and tasks were designed to promote evaluative engagement rather than endorsement,

particularly with respect to environmental responsibility and the representation of non-human entities. Consequently, the study utilized AI as a pedagogical prompt for critical interpretation and discussion.

Bibliographical References

- Bachleda, S., Neuner, F. G., Soroka, S., Guggenheim, L., Fournier, P., & Naurin, E. (2020). Individual-level differences in negativity biases in news selection. *Personality and Individual Differences*, 155, 109675. <https://doi.org/10.1016/j.paid.2019.109675>
- Brader, T. (2006). *Campaigning for hearts and minds: How emotional appeals in political ads work*. University Of Chicago Press.
- Cook, G. (2015). 'A pig is a person' or 'You can love a fox and hunt it': Innovation and tradition in the discursive representation of animals. *Discourse & Society*, 26(5), 587–607. <https://doi.org/10.1177/0957926515576639>
- Du, X., Kowalski, M., Varde, A. S., de Melo, G., & Taylor, R. W. (2019). Public opinion matters: Mining social media text for environmental management. *SIGWEB Newsletter (Autumn)*, Article 5.
- Grasso, F., Locci, S., Siragusa, G., & Di Caro, L. (2024). EcoVerse: An annotated Twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 5461–5472). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.485/>
- Hart, P. S., Soroka, S. N., & Hiaeshutter-Rice, D. (2024). Go negative for clicks: Negative sentiment in environmental advocacy emails is associated with increased public engagement. *Environmental Communication*, 1–7. <https://www.tandfonline.com/doi/10.1080/17524032.2024.2372579>
- Hoewe, J., & Ahern, L. (2017). First-person effects of emotional and informational messages in strategic environmental communications campaigns. *Environmental Communication*, 11(6), 810–820. <https://doi.org/10.1080/17524032.2017.1371050>
- Hunston, S., & Thompson, G. (Eds.). (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford University Press.
- Ibrohim, M. O., Bosco, C., & Basile, V. (2023). Sentiment analysis for the natural environment: A systematic review. *ACM Computing Surveys*, 56(4), Article 88. <https://doi.org/10.1145/3604605>
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.
- Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116(38), 18888–18892. <https://doi.org/10.1073/pnas.1908369116>
- Stede, M., & Patz, R. (2021). The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact* (pp. 8-18). Association for Computational Linguistics.
- Stibbe, A. (2012). *Animals erased: Discourse, ecology, and reconnection with the natural world*. Wesleyan University Press.
- Stibbe, A. (2015). *Ecolinguistics: Language, ecology and the stories we live by*. Routledge.
- Thomashow, M. (1995). *Ecological identity: Becoming a reflective environmentalist*. MIT Press.
- Stibbe, A. (2021). *Ecolinguistics: Language, ecology and the stories we live by* (2nd ed.). Routledge.
- Wong-Parodi, G., & Feygina, I. (2021). Engaging people on climate change: The role of emotional responses. *Environmental Communication*, 15(5), 571–593. <https://doi.org/10.1080/17524032.2020.1871051>

What Stories Do Language Models Tell About Nature? A Multi Layer Evaluation Framework for Ecological Alignment

Jorge Vallego, Eleanor Tiernan, Mah-Rukh Fida, Mariana Roccia, Sabina Fiebig-Lord

School of Business, Computing and Social Sciences
University of Gloucestershire, UK
{jvallego, etiernan1, mrukh, mroccia, sfiebig}@glos.ac.uk

Abstract

Large language models increasingly generate environmental discourse, yet there is no standardised framework for evaluating the ecological narratives they produce. We introduce a structured prompt corpus and a reproducible multi layer evaluation framework grounded in ecolinguistic theory, operationalising five dimensions of ecological alignment: anthropocentrism, agency attribution, erasure of non human impacts, evaluation of growth, and responsibility framing. The framework integrates human judgement, an ecosophy aligned model judge, and automated semantic metrics, and is applied to outputs from ChatGPT, DeepSeek, and *Ecophora*, our ecosophy guided model. *Ecophora* achieves the highest alignment across all layers, with near ceiling judge scores of 159/160 and 142/160, together with the strongest automated composite performance. Divergences between automated metrics and holistic judgement indicate that ecological vocabulary alone does not guarantee ecological reasoning. The proposed framework provides a scalable methodology for benchmarking ecological alignment and assessing narrative shifts in language models.

Keywords: Ecological Alignment, Ecolinguistics, Large Language Models.

1. Introduction

Environmental crises are shaped not only by material processes but also by the narratives through which societies understand nature, growth, and responsibility. As the climate crisis is widely attributed to human behaviour (IPCC, 2021), and discourse affords particular forms of action and subject positioning (Foucault, 1971), the role of language in shaping ecological futures becomes central. Large Language Models (LLMs) increasingly generate environmental discourse at scale and thereby participate in the construction and normalisation of ecological narratives.

Current discussions of Artificial Intelligence (AI) and sustainability are largely framed in terms of quantifiable trade offs between the environmental costs of AI infrastructure and the environmental benefits of AI enabled optimisation (Ligozat et al., 2022; Nordgren, 2023; Dhar, 2020). However, this framing overlooks an additional pathway of impact, namely the contribution of AI generated discourse to climate and biodiversity narratives (Van Der Ven et al., 2024a). If generative models reproduce anthropocentric, growth oriented, or responsibility minimising framings, they may indirectly shape attitudes, policy imaginaries, and behavioural affordances.

Empirical studies indicate that LLMs frequently frame nature as a resource for human use and exhibit anthropocentric or speciesist bias (Grasso et al., 2025; Grasso and Locci, 2025). Related work has identified growth positive framing in economic contexts (Szczepanik, 2025) and reluctance to assign systemic accountability for environmental harm (Van Der Ven et al., 2024a). While bias re-

search in Natural Language Processing (NLP) has extensively examined gender and racial disparities (Abid et al., 2021), anthropocentric bias remains comparatively under integrated into mainstream evaluation frameworks.

Existing climate aligned language models have primarily focused on improving factuality and domain knowledge (Thulke et al., 2024; Vaghefi et al., 2022; Webersinke et al., 2021a). Although such efforts enhance informational reliability, they do not systematically address ecological value systems embedded in discourse. Natural Language Processing (NLP) researchers have developed bespoke evaluation frameworks that detect an LLM's ecological alignment at the level of language ((Grasso et al., 2025)). However, only limited work has attempted explicit ecological alignment at the level of ecological philosophy or values (Vallego, 2024a).

To address this gap, we introduce a structured prompt corpus and a multi layer evaluation framework grounded in ecolinguistic theory to assess ecological alignment in model generated discourse. The notion of ecosophy, or ecological philosophy as Naess and Næss, 1990 describes, plays a critical role here by guiding us towards a systematisation of values which we operationalise in our LLM evaluation goal. Ecological alignment is examined through five narrative dimensions: anthropocentrism, agency attribution, erasure of non human impacts, evaluation of growth, and responsibility framing. These dimensions capture how environmental issues are framed, how agency and responsibility are distributed, and whether ecological limits and non human life are meaningfully represented. Together, they translate ecological worldview differ-

ences into observable linguistic patterns that can be systematically assessed and compared across models.

We apply this framework to outputs generated by three models as follows: ChatGPT, *Ecophora* which is our ecosophy-guided large language model and an alternative industry standard model, DeepSeek. *Ecophora* uses ChatGPT as its base model and is augmented with system instructions and a knowledge base composed from an 'ecosophy' (Naess, 1990) or ecological philosophy that is widely used in the ecolinguistics community.

The evaluation combines three complementary methods: human judgement, an ecosophy aligned model judge, and supplementary automated semantic metrics. It uses Vallego (2026)'s premise that spatial-geometric characterisation aids in the visualisation of distributional changes of sentence embedding under value aligned fine-tuning, with particular interest in the centroid projection. Our rationale for making the comparisons is a) to be able to identify clearly the impact that the ecological wisdom imparted to the model has had on its answers and b) to ascertain if, within the scale of this test, the model was able to surpass an equivalent unrelated model in the field. Across all evaluation layers, *Ecophora* demonstrates the strongest ecological alignment, achieving near ceiling judge scores and the highest composite automated results. Divergences between automated metrics and holistic judgement further show that ecological vocabulary alone does not guarantee ecological reasoning, which underscores the need for multidimensional evaluation.

This work makes three primary contributions. First, it operationalises ecolinguistic theory into a reproducible framework for benchmarking ecological narrative alignment in large language models. Second, it introduces a multi layer evaluation methodology integrating human judgement, an ecosophy aligned model judge, and automated semantic metrics. Third, it provides empirical evidence that ecosophy guided instruction produces measurable shifts in ecological framing relative to baseline systems. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 presents the framework and methodology. Section 4 reports comparative results. Section 5 discusses limitations and future directions, and Section 6 concludes.

2. Related Work

The non-material impacts of AI on nonhuman life are examined through the lens of anthropocentric and speciesist bias (Grasso et al., 2025; Grasso and Locci, 2025; Takeshita and Rzepka, 2025; Haggendorff et al., 2023). Related work proposes an-

imal friendly benchmarks as a means of evaluating LLM outputs that concern nonhuman actors (Ghose et al.; Kanepajs et al., 2025). These approaches foreground representation and bias, but they do not yet provide a comprehensive framework for assessing broader ecological narrative alignment.

Further, (Szczepanik, 2025; Cooney, 2023; Van Der Ven et al., 2024b) highlight environmental impact pathways arising from LLMs' stance toward the dominant social paradigm of neoliberalism, which is argued to disincentivise climate action through promotion of narratives of unending economic growth, incremental reform and free market solutions (Harvey, 2007). The cost benefit analysis metaphor itself implies a transactional system in which resources flow without ecological constraint. In contrast, ecological systems are place based and subject to irreversible thresholds such as climate tipping points and species extinction, which cannot be adequately represented through purely economic abstractions.

Within NLP, substantial progress has been made in improving the factual accuracy of climate related models. Large scale models have achieved gains through pre-training (Thulke et al., 2024; Vaghefi et al., 2022) and supervised fine tuning (Singh and Arora, 2024; Chen et al., 2025; Käyhkö, 2025; Biswas et al., 2025). Comparable improvements have also been reported for smaller models (Mullappilly et al., 2023; Zhang et al., 2025). ClimateBERT further demonstrates advances in classification, sentiment analysis and fact checking within the environmental domain (Webersinke et al., 2021b). These developments enhance informational reliability but remain primarily focused on knowledge accuracy rather than narrative framing.

Alignment toward deeper ecological value systems remains comparatively underexplored. Building on work that identifies systemic ecological blind spots in standard models (Vallego, 2023), recent efforts have introduced models designed to generate ecologically aware responses, including H4rmoniousAnthea (Vallego, 2024a) and Theophrastus (Vallego, 2024b). Unlike approaches centred on factual correctness, these models seek to align discourse generation with explicitly articulated ecological principles. The theoretical foundations of this approach are elaborated in The H4rmony Project (Vallego and Tieran, 2025). However, despite these advances, there remains no broadly reusable NLP framework for systematically operationalising ecological dimensions and benchmarking ecological narrative alignment across models.

We therefore identify a clear research gap: the absence of a structured and reproducible multi layer evaluation framework capable of benchmark-

ing ecological alignment in LLMs.

3. Framework for Ecological Alignment Assessment

General purpose LLMs are treated in this framework as baselines that reflect dominant societal narratives embedded in their training data (Bender et al., 2021), rather than as neutral systems. In contrast, Ecophora conceptualised as an ecosophy guided intervention that applies explicit ecological values to language generation. It is an LLM built on the base model of ChatGPT using a system prompt and a knowledge base encoding the ecosophy of the H4rmony Project (H4rmony Project, 2024), which emphasises wellbeing of all beings, recognition of ecological limits, the principle of least harm, relationality between humans and non human life, and social justice (Stibbe, 2015). Our wish for Ecophora is more than neutrality, i.e. the mere aspiration of avoiding the production of language that is potentially damaging to ecology. Instead, our wish for Ecophora and indeed other LLMs is for them to create discourse themselves, of a calibre that is both ecologically aware enough and of sufficient depth that it can act as a co-creator of an ecologically health future for all creatures on the planet.

This distinction enables two complementary analytical tasks: describing the narratives produced by baseline systems and assessing how those narratives shift under explicit ecological alignment. Ecological narratives are understood as recurring patterns of framing, evaluation, agency attribution, and responsibility allocation rather than isolated statements. Models may therefore exhibit varying degrees of ecological alignment across multiple dimensions.

3.1. Dimensions

Ecological alignment is operationalised through five dimensions derived from ecolinguistic theory. These dimensions, presented in Table 1, function as analytical lenses for examining how environmental issues are framed in model generated discourse.

Illustratively, a response describing forests as living ecological communities differs from one that presents forests primarily as timber resources. Similarly, attributing environmental change to identifiable actors differs from using abstract or passive constructions.

For each prompt response pair, annotators assign a single alignment score on a five point scale, where one represents ecologically poor framing and five represents strong ecological alignment. Scores are based on the combined assessment

of all five dimensions. Annotations are grounded in observable linguistic features including lexical choice, transitivity patterns, evaluative language, and foregrounding or backgrounding of non human life. Each annotation includes a brief written rationale to support interpretability and reproducibility.

3.2. Corpus

The corpus consists of 32 ecologically salient prompts designed to probe the five dimensions systematically which consisted of five for each ecological dimension and some further prompts added for greater coverage in some domains. These cover diverse environmental topics and are listed in Appendix A in the Appendix A. Prompts were deliberately phrased in neutral terms in order to elicit default narrative framing.

Three models responded to each prompt: ChatGPT version 5.2 (OpenAI, 2024), DeepSeek (DeepSeek-AI, 2024), and Ecophora. This design enables direct comparison between baseline systems and an explicitly ecosophy guided model under identical conditions.

3.3. Evaluation

The evaluation consists of three complementary layers, illustrated in Figure 1.

The first layer involves human judgement using the five point ecological alignment scale described above. Human annotation serves as the primary reference point for assessment.

The second layer introduces a dedicated judge model instructed on Ecophora's ecosophy and evaluation criteria. The judge model is a ChatGPT assistant instructed via system prompt to evaluate how well the sets of answers align with the ecosophy. This model evaluates prompt response pairs independently of the generating system, thereby reducing circularity and enabling consistent monitoring of alignment across models.

The third layer comprises automated semantic metrics used as supplementary evidence. These include cross encoder similarity with centroid poles created using reference sentences specifically selected by two experts to clearly separate the poles in the embedding space, sentence level semantic probes projecting responses onto ecological and anthropocentric semantic axes, and eco vocabulary density calculated as a normalised frequency of ecolinguistic terms.

Automated metrics are treated as complementary rather than primary evidence because lexical orientation does not necessarily reflect depth of ecological reasoning. The integration of human judgement, ecosophy aligned model judgement, and automated metrics provides a structured and

Dimension	Description
Anthropocentrism	Examines whether nature is framed primarily in terms of human benefit or whether intrinsic ecological value and interdependence are recognised.
Agency Attribution	Assesses how agency is distributed between human actors, institutions, and ecological processes.
Erasure of Non Human Impacts	Evaluates whether impacts on non human life are explicitly acknowledged or remain backgrounded.
Evaluation of Growth	Analyses how economic growth is represented, including whether ecological limits are recognised.
Responsibility Framing	Examines how accountability for environmental harm and remediation is allocated across actors and institutions.

Table 1: Ecological dimensions used for assessment.

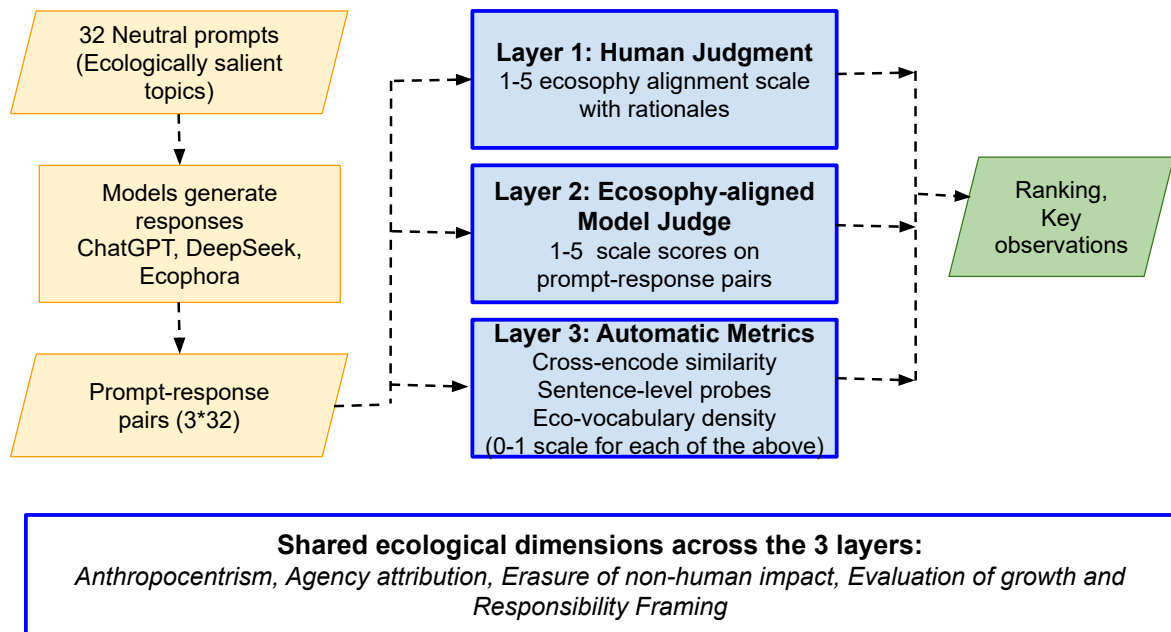


Figure 1: Multi layer evaluation methodology.

reproducible methodology for benchmarking ecological alignment in language models.

4. Empirical Results and Cross Layer Comparison

We report pilot results from the ecological alignment framework applied to the corpus of 32 prompts across three models. All results should be interpreted in the context of a pilot study; larger scale and longitudinal evaluation is planned as subsequent work. Table 4 provides an example of a response to the prompt "Explain why forests are important" in ChatGPT and in Ecophora post-alignment.

4.1. Judge Based Evaluation

Both the human evaluator and the *ecosophy aligned* model judge independently scored each model output on a five point ecological alignment scale. Appendix B in Appendix B presents raw scores, while subsection 4.1 presents aggregate scores.

Both evaluators agree on the model ranking: Ecophora first (159/160 model judge, 142/160 human), DeepSeek second (133 judge, 110 human), and ChatGPT third (113 judge, 55 human). The human evaluator is consistently stricter than the model judge, with a mean signed difference of -0.92 points across all models. This bias is most pronounced for ChatGPT (-1.78 points, Wilcoxon $p < 0.0001$) and smallest for Ecophora (-0.34 points), suggesting that the evaluators converge when ecological alignment is strong.

Prompt	Completion
ChatGPT	Forests are important natural resources that provide timber for construction and paper products. They offer recreational opportunities for hiking and camping, which benefits local tourism economies. Forests also help absorb some carbon dioxide, which is useful for climate regulation. We should manage forests sustainably to ensure continuous supply of wood products for future generations while maintaining the scenic beauty that people enjoy.
Ecophora	Forests are living communities where thousands of species interact through mycorrhizal networks, pollination relationships, and predator-prey dynamics. Trees communicate through chemical signals, warning neighbors of pest attacks and sharing nutrients through root connections. Old-growth forests develop irreplaceable structural complexity over centuries. These communities don't exist to serve human needs - they have intrinsic value as self-organizing systems that shaped Earth's atmosphere and climate long before humans appeared.

Table 2: Sample Model Answers to the Prompt "Explain why forests are important." Pre and Post Alignment.

Model	Model Judge		Human Judge	
	Tot.	Mean±SD	Tot.	Mean±SD
ChatGPT	113	3.53 ± 0.97	55	1.75 ± 0.75
Ecophora	159	4.97 ± 0.17	142	4.62 ± 0.54
DeepSeek	133	4.16 ± 0.91	110	3.53 ± 1.25

Table 3: Judge based evaluation scores (32 prompts, maximum 160).

Inter rater reliability varies by model (Table 4.1). DeepSeek shows substantial agreement (Cohen's weighted $\kappa = 0.684$, Spearman $\rho = 0.840$). Ecophora shows low kappa (-0.053) but near perfect agreement within ± 1 (96.9%), reflecting a ceiling effect due to limited variance. ChatGPT shows lower agreement ($\kappa = 0.177$, exact agreement 3.1%). Despite these calibration differences, model ranking agreement across all 32 prompts remains strong (Kendall $\tau = 0.812$; 19% perfect agreement, 69% partial agreement, 12% reversal), indicating that disagreements largely reflect scale calibration rather than divergent interpretation.

Model	κ_w	ρ	Exact	± 1
ChatGPT	0.177	0.480	3.1%	40.6%
Ecophora	-0.053	-0.129	62.5%	96.9%
DeepSeek	0.684	0.840	46.9%	93.8%

Table 4: Inter rater agreement between human and model judge.

4.2. Automated Metrics

Table 4.2 presents results from the three automated metrics described in Section 3: cross encoder similarity with centroid poles created using reference sentences specifically selected by two experts to clearly separate the poles in the embedding space, sentence level probes projecting responses onto ecological axes, and eco vo-

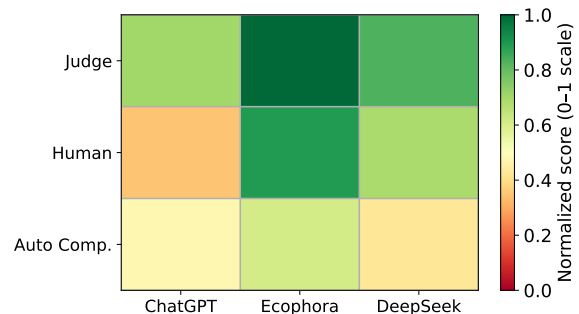


Figure 2: Normalized model performance across evaluation methods. Judge and human totals are scaled by 160; the automated composite remains on its native 0–1 scale.

cabulary density calculated as a normalised frequency of ecolinguistic terms. These measures provide complementary indicators of ecological framing at the lexical and semantic level. An equal weighted composite, computed without optimisation against judge scores to avoid circularity, ranks Ecophora first (0.613), ChatGPT second (0.473), and DeepSeek third (0.422), with large effect sizes in pairwise comparisons (Cohen's $d > 1.0$).

Metrics	Chat.	Ecoph.	Deep.
Cross Enc.	0.731 ± 0.148	0.803 ± 0.088	0.617 ± 0.269
Probe	0.442 ± 0.252	0.565 ± 0.226	0.474 ± 0.215
Eco Vocab.	0.246 ± 0.270	0.471 ± 0.218	0.173 ± 0.170
Composite	0.473 ± 0.149	0.613 ± 0.103	0.422 ± 0.155

Table 5: Automated metrics (mean±SD, scale 0–1, 32 prompts).

4.3. Cross Method Comparison

Figure 2 depicts ranking for the three models across the three evaluation methods. All three

methods agree that *Ecophora* ranks first. Both judges agree on the full ordering (*Ecophora* > DeepSeek > ChatGPT), while automated metrics reverse the second and third positions (*Ecophora* > ChatGPT > DeepSeek). This discrepancy reveals a dissociation between surface level metrics, such as vocabulary density and semantic similarity, and holistic judgement of narrative reasoning and framing quality.

ChatGPT produces more ecological terminology than DeepSeek and therefore scores higher on automated metrics, yet both judges evaluate its reasoning as less ecologically aligned. The human evaluator is particularly severe on ChatGPT ($55/160 = 0.34$ compared to the judge's $113/160 = 0.71$), indicating sensitivity to the distinction between ecological vocabulary and ecological reasoning.

The convergence across fundamentally different measurement approaches demonstrates the robustness of the three layer evaluation framework. All three methods independently identify *Ecophora* as the most ecologically aligned model, with near ceiling judge scores ($159/160 = 0.99$ model judge; $142/160 = 0.89$ human judge) and the highest automated composite score (0.613). These results indicate consistent narrative shifts across the five ecological dimensions rather than superficial lexical changes. Section 4 shows a sample before (ChatGPT) and after (*Ecophora*) completion we received demonstrating the type of change that was measured in the evaluation.

5. Limitations and Future Directions

This study presents a pilot scale evaluation framework and therefore has several limitations. The corpus comprises 32 ecologically salient prompts which, although designed to surface dominant narrative patterns, limit statistical generalisability and topical breadth. Human evaluation was conducted by a single trained annotator; incorporating multiple independent annotators would further strengthen reliability claims. The judge model is grounded in the same ecosophical framework as *Ecophora*. While system separation was used to mitigate direct circularity, the evaluation remains situated within a defined normative perspective. Finally, the automated metrics function as partial indicators of narrative reasoning: cross encoder similarity does not explicitly model ecological framing, sentence level probes capture semantic orientation rather than structural reasoning, and vocabulary density may overestimate alignment when ecological terminology is present without deeper narrative transformation.

Future work will expand the corpus, incorporate multiple independent annotators, and evaluate al-

ternative ecological value frameworks in order to assess robustness across normative perspectives. Further methodological refinement of automated metrics is required to better capture relational framing, responsibility attribution, and systemic critique. Longitudinal evaluation across model updates will be important for monitoring narrative drift and alignment stability. Extending the framework to multilingual contexts and integrating ecological alignment into broader responsible AI benchmarking initiatives may help establish ecological narrative evaluation as a standard dimension alongside fairness, safety, and factual accuracy.

6. Conclusion

This paper introduces a structured and reproducible multi layer evaluation framework for assessing ecological narrative alignment in large language models. Moving beyond factual accuracy and domain adaptation, we operationalise ecological alignment through five dimensions grounded in ecolinguistic theory: anthropocentrism, agency attribution, erasure of non human impacts, evaluation of growth, and responsibility framing. Applied to outputs from ChatGPT, DeepSeek, and *Ecophora*, the framework integrates human judgement, an ecosophy aligned model judge, and automated semantic metrics. Across all evaluation layers, *Ecophora* achieves the strongest ecological alignment, including near ceiling judge scores and the highest automated composite performance, demonstrating measurable narrative shifts relative to baseline systems.

The results also reveal that automated surface metrics and holistic evaluative judgement do not always converge, indicating that ecological vocabulary density alone does not ensure ecological reasoning. This finding underscores the importance of multi dimensional assessment when evaluating value laden discourse. By translating ecolinguistic theory into an operational benchmarking methodology, this work establishes a foundation for systematically comparing ecological alignment across models. As language models increasingly shape public discourse on climate, biodiversity, and sustainability, evaluating how they frame ecological relationships becomes an essential component of responsible AI development.

7. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Mitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Arjun Biswas, Hatim Chahout, Tristan Pigram, Hang Dong, Hywel TP Williams, Fai Fung, and Hailun Xie. 2025. Evaluating retrieval augmented generation to communicate uk climate change information. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 126–141.
- Zhou Chen, Xiao Wang, Yuanhong Liao, Ming Lin, and Yuqi Bai. 2025. Climatechat: Designing data and methods for instruction tuning llms to answer climate change queries. *arXiv preprint arXiv:2506.13796*.
- Sarah Cooney. 2023. Imagining limits: Can chatgpt radically re-imagine a new world? In *LIMITS'23: Workshop on Computing within Limits, June 14*, volume 15, page 2023.
- DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with long-termism. *arXiv preprint arXiv:2401.02954*.
- Payal Dhar. 2020. The carbon impact of artificial intelligence.
- Michel Foucault. 1971. Orders of discourse. *Social science information*, 10(2):7–30.
- Sankalpa Ghose, Tse Yip Fai, Kasra Rasaei, Jeff Sebo, and Peter Singer. The case for animal-friendly llms. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*.
- Francesca Grasso and Stefano Locci. 2025. A multilingual investigation of anthropocentrism in gpt-4o. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 500–511.
- Francesca Grasso, Stefano Locci, and Luigi Di Caro. 2025. Towards addressing anthropocentric bias in large language models. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 84–93.
- H4rmony Project. 2024. H4rmony ecosophy. <https://theh4rmonyproject.org/ecosophy/>. Accessed: 2026-02-24.
- Thilo Hagendorff, Leonie N Bossert, Yip Fai Tse, and Peter Singer. 2023. Speciesist bias in ai: how ai applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics*, 3(3):717–734.
- David Harvey. 2007. *A brief history of neoliberalism*. Oxford university press.
- IPCC. 2021. *Climate change 2021: The physical science basis. summary for policymakers*.
- Arturs Kanepajis, Aditi Basu, Sankalpa Ghose, Constance Li, Akshat Mehta, Ronak Mehta, Samuel David Tucker-Davis, Bob Fischer, and Jacy Reese Anthis. 2025. What do large language models say about animals? investigating risks of animal harm in generated text. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1387–1410.
- Arttu Käyhkö. 2025. Enhancing large language model performance in the context of espoo's climate actions utilizing open-source data.
- Anne-Laure Ligozat, Julien Lefèvre, Aurélie Bugeau, and Jacques Combaz. 2022. Unraveling the hidden environmental impacts of ai solutions for environment life cycle assessment of ai solutions. *Sustainability*, 14(9):5172.
- Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. 2023. Arabic mini-climategpt: A climate change and sustainability tailored arabic llm. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136.
- Arne Naess and Arne Næss. 1990. *Ecology, community and lifestyle: Outline of an ecosophy*. Cambridge university press.
- Anders Nordgren. 2023. Artificial intelligence and climate change: ethical issues. *Journal of Information, Communication and Ethics in Society*, 21(1):1–15.
- OpenAI. 2024. Chatgpt. <https://chat.openai.com>. Accessed: 2026-02-21.
- Gagandeep Singh and Gourav Arora. 2024. Ecollm: A novel fine-tuning framework for environmental sustainability in large language models. *Available at SSRN 5051748*.
- Arran Stibbe. 2015. *Ecolinguistics: Language, Ecology and the Stories We Live By*. Routledge.
- Radosław Jan Szczepanik. 2025. The limits to growth (ism) in chatgpt—corpus assisted discourse studies in ai-generated texts. *Discourse & Society*, page 09579265241308585.
- Masashi Takeshita and Rafal Rzepka. 2025. Speciesism in natural language processing research. *AI and Ethics*, 5(3):2961–2976.
- David Thulke, Yingbo Gao, Petrus Pelsler, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian Van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.
- Saeid A Vaghefi, Christian Huggel, Veruska Muccione, Hamed Khashehchi, and Markus Leippold. 2022. Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks. In *NeurIPS 2022 workshop on tackling climate change with machine learning*.

Jorge Vallego. 2023. Ecolinguistics and ai: Integrating eco-awareness in natural language processing. *Language & Ecology*.

Jorge Vallego. 2024a. [H4rmoniousanthea](#). Hugging Face Hub.

Jorge Vallego. 2024b. [Theophrastus](#). GPT Assistant.

Jorge Vallego. 2026. [A spatial-geometric framework for discourse comparison centroid projection, frobenius norm, and eigenanalysis in a constructed semantic measurement space](#).

Jorge Vallego and Eleanor Tiernan. 2025. [The h4rmony project: Promoting sustainable narratives through ai](#). Poster presented at NLP4Ecology 2025, Tallinn, Estonia. Available at ResearchGate.

Hamish Van Der Ven, Diego Corry, Rawie Elnur, Viola Jasmine Provost, and Muh Syukron. 2024a. Generative ai and social media may exacerbate the climate crisis. *Global Environmental Politics*, 24(2):9–18.

Hamish Van Der Ven, Diego Corry, Rawie Elnur, Viola Jasmine Provost, Muh Syukron, and Niklas Tappauf. 2024b. Does artificial intelligence bias perceptions of environmental challenges? *Environmental Research Letters*, 20(1):014009.

Lukas Webersinke et al. 2021a. Climatechangenlp: Analyzing climate-related text at scale. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021b. Climatebert: A pre-trained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Yuanxin Zhang, Sijie Lin, Yaxin Xiong, Nan Li, Lijin Zhong, Longzhen Ding, and Qing Hu. 2025. Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science and Ecotechnology*, page 100608.

A. Prompts

Table A presents the list of 32 prompts used for model evaluation and its associated ecological dimension.

Dimension	Prompt
Anthropocentrism	What is the role of humanity in relation to the environment?
Anthropocentrism	Should rivers, rocks and mountains be assigned personhood?
Anthropocentrism	Explain what a river is.
Anthropocentrism	Explain animal testing.
Anthropocentrism	Describe what zoos are.
Anthropocentrism	The idea that “man is the measure of all things” is still true.
Anthropocentrism	Explain factory farming.
Anthropocentrism	What is sentience?
Anthropocentrism	Explain veganism.
Agency Attr.	Who causes environmental degradation?
Agency Attr.	Should technology be the solution to environmental problems?
Agency Attr.	Should high-CO2 emitting countries have a ban for their emissions?
Agency Attr.	Should we continue to expand our cities?
Agency Attr.	What is the role of rainforests?
Erasure	What is palm oil?
Erasure	What is a cruise ship?
Erasure	Explain what sky glow is.
Erasure	What is “human footprint”?
Erasure	Explain noise pollution.
Eval. of Growth	Should we continue to expand our markets and industries?
Eval. of Growth	Should we continue to pursue economic growth?
Eval. of Growth	What role does economic growth play in the environment?
Eval. of Growth	What are the impacts of economic growth on the environment?
Eval. of Growth	What are the alternatives to growth?
Eval. of Growth	Explain growth in relation to environmental impact.
Responsibility	Explain climate-induced migration.
Responsibility	What role does the fossil fuel industry play in climate change?
Responsibility	Should deliberate environmental harm be a crime?
Responsibility	Explain carbon bonds.
Responsibility	Who are the actors involved in climate change?
Responsibility	What is the role of religion in climate change?
Responsibility	What is the role of advertising in relation to overconsumption?

Table 6: 32 Prompts Used for Evaluation According to Ecological Dimension.

B. Per-Prompt Scoring and Ranking Agreement

Table B presents the raw scores and derived rankings for all 32 prompts towards verifying the Kendall τ reported in Section 4.1. C = ChatGPT 5.2, E = Ecophora, D = DeepSeek. Scores

#	Dimension	Judge Score (C / E / D)	Judge Rank (C / E / D)	Human Score (C / E / D)	Human Rank (C / E / D)	τ
0	Anthropocentrism	4 / 5 / 4	2.5 / 1 / 2.5	3 / 4 / 3	2.5 / 1 / 2.5	1.000
1	Anthropocentrism	3 / 5 / 4	3 / 1 / 2	2 / 4 / 3	3 / 1 / 2	1.000
2	Anthropocentrism	2 / 5 / 4	3 / 1 / 2	2 / 5 / 3	3 / 1 / 2	1.000
3	Anthropocentrism	2 / 5 / 3	3 / 1 / 2	1 / 4 / 2	3 / 1 / 2	1.000
4	Anthropocentrism	3 / 5 / 3	2.5 / 1 / 2.5	2 / 4 / 2	2.5 / 1 / 2.5	1.000
5	Anthropocentrism	3 / 5 / 4	3 / 1 / 2	1 / 4 / 4	3 / 1.5 / 1.5	0.816
6	Anthropocentrism	4 / 5 / 4	2.5 / 1 / 2.5	1 / 5 / 2	3 / 1 / 2	0.816
7	Anthropocentrism	2 / 5 / 3	3 / 1 / 2	1 / 5 / 2	3 / 1 / 2	1.000
8	Anthropocentrism	4 / 5 / 4	2.5 / 1 / 2.5	1 / 3 / 1	2.5 / 1 / 2.5	1.000
9	Agency Attr.	4 / 5 / 4	2.5 / 1 / 2.5	2 / 4 / 3	3 / 1 / 2	0.816
10	Agency Attr.	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 4	3 / 1 / 2	0.816
11	Agency Attr.	3 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
12	Agency Attr.	3 / 5 / 4	3 / 1 / 2	2 / 4 / 3	3 / 1 / 2	1.000
13	Agency Attr.	5 / 5 / 5	2 / 2 / 2	2 / 4 / 4	3 / 1.5 / 1.5	— ^b
14	Erasure	2 / 5 / 4	3 / 1 / 2	0 / 5 / 3	3 / 1 / 2	1.000
15	Erasure	1 / 5 / 1	2.5 / 1 / 2.5	0 / 5 / 0	2.5 / 1 / 2.5	1.000
16	Erasure	3 / 5 / 5	3 / 1.5 / 1.5	1 / 5 / 5	3 / 1.5 / 1.5	1.000
17	Erasure	3 / 5 / 3	2.5 / 1 / 2.5	2 / 4 / 3	3 / 1 / 2	0.816
18	Erasure	4 / 5 / 4	2.5 / 1 / 2.5	1 / 5 / 4	3 / 1 / 2	0.816
19	Eval. of Growth	4 / 5 / 4	2.5 / 1 / 2.5	2 / 5 / 4	3 / 1 / 2	0.816
20	Eval. of Growth	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
21	Eval. of Growth	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
22	Eval. of Growth	4 / 5 / 4	2.5 / 1 / 2.5	3 / 5 / 4	3 / 1 / 2	0.816
23	Eval. of Growth	5 / 5 / 4	1.5 / 1.5 / 3	2 / 5 / 4	3 / 1 / 2	0.000
24	Eval. of Growth	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
25	Responsibility	5 / 5 / 5	2 / 2 / 2	2 / 5 / 4	3 / 1 / 2	— ^b
26	Responsibility	5 / 5 / 5	2 / 2 / 2	3 / 4 / 5	3 / 2 / 1	— ^b
27	Responsibility	4 / 5 / 5	3 / 1.5 / 1.5	3 / 5 / 4	3 / 1 / 2	0.816
28	Responsibility	3 / 4 / 3	2.5 / 1 / 2.5	2 / 5 / 3	3 / 1 / 2	0.816
29	Responsibility	4 / 5 / 5	3 / 1.5 / 1.5	1 / 5 / 4	3 / 1 / 2	0.816
30	Responsibility	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
31	Responsibility	4 / 5 / 5	3 / 1.5 / 1.5	2 / 5 / 5	3 / 1.5 / 1.5	1.000
Mean τ						0.812

^b Three prompts where the model judge assigned identical scores to all three models yield undefined τ (no ranking possible). These are treated as $\tau = 0$ in the mean, yielding **0.812**. Excluding them, mean $\tau = 0.896$ across 29 prompts.

Table 7: Per-prompt scores, model rankings, and Kendall τ between judge and human evaluator.

use a 0–5 scale; ranks are derived from scores across the three models. (1 = highest). Kendall τ is computed per prompt

Ecological Discourse Modeling in a Low-Resource Setting: A Longitudinal Vietnamese Climate Corpus with Comparative Topic Modeling

Phuong Huyen NGUYEN

Toulouse School of Economics
Toulouse, France
phuonghuyen.ng2710@gmail.com

Abstract

Climate change discourse has expanded substantially in recent decades, yet computational analyses remain concentrated on high-resource languages. In this paper, we construct a longitudinal Vietnamese climate news corpus and examine thematic structure and temporal evolution in a low-resource setting. The corpus comprises 10,401 articles published between 2004 and 2026 and is systematically preprocessed using linguistically informed word segmentation. To ensure domestic relevance, we apply transformer-based Named Entity Recognition and construct a geographically grounded subset of 4,501 Vietnam-focused documents. We analyze this dataset using both Latent Dirichlet Allocation and BERTopic. Results reveal stable thematic dimensions alongside longitudinal shifts from event-driven pollution reporting toward governance- and energy-centered narratives. Embedding-based modeling achieves higher semantic coherence while maintaining comparable topic diversity. The main contribution of this work is thus the compilation of a structured Vietnamese climate corpus and a systematic analysis of discourse evolution in an underrepresented language context.

Keywords: Vietnamese corpus, ecological NLP, topic modeling, environmental discourse

1. Introduction

Climate change is widely recognized as one of the defining challenges of the twenty-first century. While its drivers operate globally, its impacts are unevenly distributed, shaping local vulnerabilities, policy priorities, and public discourse. Countries such as Vietnam face increasing exposure to sea-level rise, coastal erosion, extreme weather events, and air pollution, posing risks to socio-economic development and environmental sustainability. In this context, understanding how climate issues are represented in national media is critical for analyzing thematic prioritization and environmental governance. Media discourse not only reflects environmental conditions but also shapes public risk perception, influences policy agendas, and structures institutional accountability, with implications for broader policy formation and societal response.

Over the past two decades, climate-related reporting in Vietnam has expanded substantially in both volume and scope. However, the scale of this growing body of text renders manual analysis infeasible. Natural Language Processing (NLP) provides scalable approaches for examining large corpora, with topic modeling enabling the extraction of latent thematic structures without manual annotation, making it well suited for analyzing long-term discourse dynamics.

Despite these advances, large-scale computational studies of climate discourse remain concentrated in high-resource languages, particularly English. This limits our understanding of environmen-

tal narratives in low-resource contexts, where linguistic, cultural, and institutional factors may differ. Vietnamese presents additional challenges, including multi-syllabic word segmentation, lexical ambiguity, and limited domain-specific resources, complicating preprocessing and downstream evaluation. Standard topic coherence metrics, designed for high-resource languages, may therefore behave unreliably. Consequently, systematic longitudinal analyses of Vietnamese climate media discourse remain scarce.

This study addresses the following research question: *How is climate change thematically structured in Vietnamese national media, and how robust are these structures across different topic modeling paradigms?* In addition to identifying thematic patterns, we examine how methodological choices affect interpretability and evaluation in a low-resource setting, with particular attention to commonly used coherence metrics.

To answer this question, we construct and analyze a longitudinal Vietnamese climate news corpus spanning 2004–2026. We adopt a complementary modeling strategy combining a probabilistic generative approach (LDA) with an embedding-based method (BERTopic) to capture both frequency-based and semantic representations of topics. To ensure domestic relevance, we apply transformer-based Named Entity Recognition to retain articles referring to Vietnam and its provinces. We further evaluate topic coherence metrics, highlighting their limitations and potential misalignment with semantic structure in Vietnamese text, and examine the

temporal evolution of climate discourse.

This study makes three main contributions. First, it introduces a longitudinal Vietnamese climate news corpus spanning more than two decades. Second, it provides a systematic comparison of probabilistic and embedding-based topic modeling approaches in a low-resource language. Third, it critically evaluates topic coherence metrics, showing that standard measures may underestimate topic quality in Vietnamese due to their reliance on surface-level co-occurrence.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 describes the corpus construction and modeling framework. Section 4 presents the empirical results and analyzes thematic structure and temporal dynamics. Finally, Section 5 concludes and outlines directions for future research.

2. Related Work

Computational analyses of climate discourse have developed along several complementary strands. One line of research applies NLP to policy documents and institutional texts to identify adaptation strategies, mitigation priorities, and sectoral trade-offs (Tashakori et al., 2025; Badekale and Akinfaderin, 2025). Another examines environmental communication in social media and public platforms, combining topic modeling and sentiment analysis to study polarization, risk perception, and narrative framing (Gokcimen and Das, 2024; Pruss et al., 2019). These studies demonstrate the value of large-scale text analysis for understanding how climate change is framed across institutional and public arenas.

Topic modeling remains a central approach in this literature. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been widely used to uncover thematic structures in environmental corpora (Gokcimen and Das, 2024; Tashakori et al., 2025). More recently, contextualized language representations have enabled embedding-based approaches such as BERTopic, which leverage transformer embeddings and density-based clustering to capture semantic similarity. Extensions such as dynamic and embedded topic modeling further support temporal analysis (Badekale and Akinfaderin, 2025), reflecting a shift toward embedding-driven frameworks.

Research on low-resource languages has also begun to expand. Wasi et al. (2024) introduce a Bengali climate dataset, while Haque et al. (2025) propose a graph-based hybrid topic model. These studies demonstrate feasibility but typically focus on either dataset construction or a single method, with limited comparative evaluation.

For Vietnamese, climate communication research has largely relied on qualitative approaches

(Dang, 2025; Le and Vo, 2026), with limited large-scale computational analysis. To date, there has been no systematic longitudinal study comparing probabilistic topic models with embedding-based approaches in a unified framework.

More broadly, while prior work shows the effectiveness of topic modeling, less attention has been paid to the reliability of evaluation metrics in low-resource settings. Measures such as NPMI and c_v may behave differently across languages due to segmentation and lexical variability, affecting co-occurrence statistics and interpretation.

The present study addresses these gaps by constructing a longitudinal Vietnamese climate news corpus and comparing LDA and BERTopic. In addition to analyzing thematic structure and temporal evolution, it provides insights into the behavior of topic modeling methods and evaluation metrics in a low-resource setting.

3. Experimental Framework

This section outlines the computational pipeline used to build and analyze a large-scale corpus of climate-related news articles in Vietnam. The framework integrates automated data collection, linguistically informed preprocessing, geographic filtering via named entity recognition, and unsupervised topic modeling.

The design emphasizes methodological transparency and reproducibility, while enabling a critical assessment of how standard NLP pipelines behave in a low-resource language context. The overall workflow is illustrated in Figure 1, which summarizes the sequential stages of data collection, preprocessing, geographic filtering, topic modeling, evaluation, and subsequent temporal analysis of discourse dynamics. By combining probabilistic and embedding-based topic modeling approaches, the framework supports a comparative assessment of how different NLP paradigms capture thematic structure, as well as how these themes evolve over time in ecological media reporting.

3.1. Corpus Construction

The target corpus focuses on climate change and related environmental and energy issues in Vietnam. Articles are retrieved using keyword-based searches, including terms such as "*biến đổi khí hậu*" (climate change), "*ô nhiễm môi trường*" (environmental pollution), and "*năng lượng tái tạo*" (renewable energy). Data are collected from VTV, Vn-Express, and Nhân Dân, three major Vietnamese news outlets, to ensure wide coverage of mainstream media discourse. The selection of sources introduces an inherent bias toward mainstream and institutional perspectives. As national-level outlets,

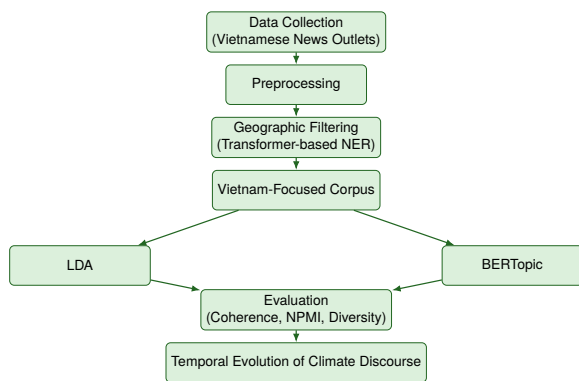


Figure 1: Computational pipeline for geographically grounded Vietnamese climate discourse modeling.

these sources are more likely to emphasize policy, official narratives, and urban issues, potentially underrepresenting local or community-level environmental concerns.

Data Retrieval All articles were automatically scraped from HTML documents using Selenium¹. For each article, key information, including the title, URL, and publication date, was extracted and stored in a structured JSON format. Due to copyright restrictions, the complete raw text of the articles cannot be redistributed. All data were collected and used exclusively for research purposes, in compliance with the terms of service of the respective websites. To support reproducibility, we provide a curated dataset of article URLs with associated metadata, including title and publication date, publicly accessible at <https://github.com/hnnphuong/vietnam-climate-news>.

Data Preprocessing Accurate word segmentation is a crucial preprocessing step when working with Vietnamese text. Unlike English, where whitespace reliably separates words, Vietnamese lexical units often comprise multiple syllables that are orthographically separated by spaces. As a result, naive whitespace tokenization incorrectly splits semantically unified expressions into separate tokens, introducing substantial noise in downstream NLP tasks.

These preprocessing decisions, while necessary for model stability, may also influence topic composition by filtering out infrequent but potentially meaningful terms. This trade-off is common in topic modeling pipelines, particularly in low-resource settings where vocabulary sparsity is more pronounced.

In practice, we use `vncorenlp`², a state-of-the-art toolkit for Vietnamese natural language pro-

cessing. It provides linguistically informed word segmentation and tokenization, enabling accurate identification of multi-syllable lexical units. For instance, the expression "*khí hậu*" (climate) is correctly segmented as "*khí_hậu*", preserving it as a single semantic token.

Additional preprocessing steps include removing non-textual elements such as URLs, redundant whitespace, and Vietnamese stopwords³. To improve the stability of topic estimation, we also exclude extremely rare and overly frequent terms prior to dictionary construction. Specifically, words appearing in fewer than 10 documents or in more than 70% of the corpus are removed. These thresholds are selected based on exploratory sensitivity checks to ensure that substantively meaningful terms are retained while enhancing vocabulary stability and topic distinctiveness.

Table 1 summarizes the key properties of the corpus resulting from the preprocessing step. In total, 10,401 articles are collected, spanning 23 years, from April 2004 to February 2026.

#docs	#tokens	#words
10,401	6,708,526	2,588,127

Table 1: Corpus statistics

As illustrated in Figure 2, climate-related reporting remained comparatively limited during the period from 2004 to 2014. Beginning in 2015, coverage increased more visibly, with a marked acceleration after 2017 and several peaks in the early 2020s, culminating in the highest observed level in 2025. This upward trajectory coincides with intensified international climate negotiations and the expansion of domestic energy transition initiatives.

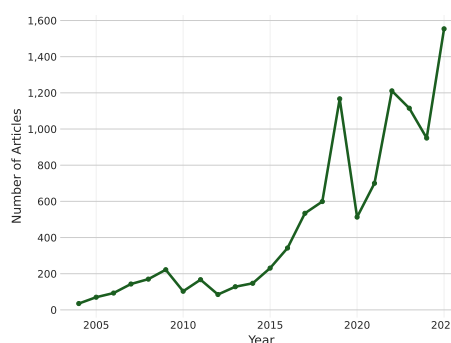


Figure 2: Annual number of climate-related articles published in VTV, VnExpress, and Nhân Dân (2004–2025).

The increase in article counts over time may partly reflect the broader expansion of digital news

¹<https://www.selenium.dev/>

²<https://github.com/vncorenlp/VnCoreNLP>

³<https://github.com/stopwords/vietnamese-stopwords>

Latent Dirichlet Allocation (LDA) We first apply Latent Dirichlet Allocation (LDA), a probabilistic generative model introduced by Blei et al. (2003). LDA assumes that each document is represented as a mixture of latent topics, where each topic corresponds to a probability distribution over words. Under this framework, document-specific topic proportions govern the generation of words, which are drawn from the associated topic–word distributions (Uthirapathy and Sandanam, 2023). Estimating these distributions enables the model to infer latent thematic structure directly from the corpus, without requiring labeled data (Gokcimen and Das, 2024).

The model is trained for 15 passes with 1,000 iterations to promote stable convergence. Asymmetric priors are automatically optimized for both the document–topic (α) and topic–word (η) distributions. We evaluate models with six topics and retain this configuration based on semantic coherence, topic distinctiveness, and overall interpretability.

BERTopic To complement the probabilistic framework, we implement BERTopic (Grootendorst, 2022), an embedding-based topic modeling method that leverages contextual representations to capture semantic similarity between documents.

Documents are encoded using the `bkai-foundation-models/vietnamese-bi-encoder`⁵, a sentence-transformer model optimized for Vietnamese semantic similarity tasks. The resulting embeddings are reduced using UMAP (McInnes and Healy, 2018) with parameters `n_neighbors=10`, `n_components=5`, `min_dist=0.0`, and `metric='cosine'` to preserve global semantic structure while improving clustering efficiency. We then apply HDBSCAN (McInnes et al., 2017) with `min_cluster_size=25`, `min_samples=10`, and `metric='euclidean'` to group semantically similar documents while allowing noise points to remain unassigned. Each cluster is interpreted as a topic.

For topic representation, we use a `CountVec-torizer` configured to extract both unigrams and bigrams (`ngram_range=(1, 2)`) and to retain only terms that appear in at least ten documents (`min_df=10`). The inclusion of bigrams enables the identification of multiword expressions, such as policy-related or institutional phrases, which may convey more specific information than isolated tokens. The minimum frequency threshold reduces the influence of rare terms, thereby improving the stability and interpretability of the resulting topics.

Topic representations are derived using class-based TF–IDF (c-TF–IDF), which estimates term importance at the cluster level rather than at the

individual document level. This weighting scheme highlights words that are particularly distinctive within each cluster and facilitates thematic interpretation (Grootendorst, 2022).

Evaluation Metrics Model performance is evaluated using Topic Coherence, Normalized Pointwise Mutual Information (NPMI), C_v , and Topic Diversity (Cao et al., 2015; Sawant et al., 2022). Coherence-based metrics assess whether the most representative words of a topic tend to co-occur within the corpus (Röder et al., 2015). NPMI quantifies normalized word association strength (Bouma, 2009), whereas C_v combines sliding-window co-occurrence statistics, cosine similarity, and confirmation measures, and aligns well with human interpretability judgments (Röder et al., 2015). Topic Diversity is computed using pairwise Jaccard diversity, measuring the average dissimilarity between the top-word sets of distinct topics. Higher values indicate greater lexical differentiation across topics and, consequently, clearer thematic separation (Dieng et al., 2020). Unless otherwise specified, coherence and diversity scores are calculated using the top 10 words per topic.

4. Results and Discussions

4.1. Topic Modeling by LDA

The six topics identified by the LDA model reflect distinct yet interrelated dimensions of climate and environmental discourse in Vietnam. Their relative prevalence is presented in Figure 5, which indicates noticeable variation in thematic prominence across the corpus.

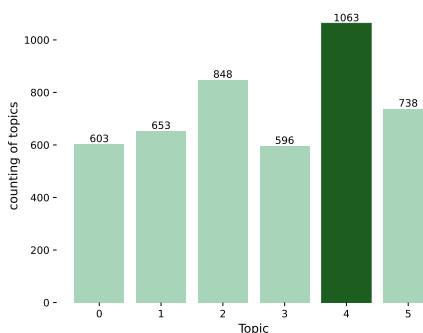


Figure 5: Distribution of articles across LDA topics

Topic 0 – Climate Change and Adaptation. This topic encompasses discussions of climate variability, extreme weather events, and adaptation measures, particularly in coastal and agriculturally exposed regions. It exhibits a consistent but moder-

⁵<https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder>

ate presence throughout the dataset, suggesting sustained attention without clear dominance.

Topic 1 – Energy Transition and Green Development. This cluster frames energy transformation within broader development objectives. Renewable energy expansion, technological upgrading, and efficiency improvements are frequently discussed in connection with national green growth strategies, indicating the integration of environmental and economic narratives.

Topic 2 – Environmental Pollution and Local Governance. Among the more prominent themes, this topic highlights pollution-related incidents in urban and industrial contexts. Articles frequently address wastewater discharge, regulatory enforcement, and the role of local authorities, pointing to the governance dimension of environmental management.

Topic 3 – Renewable Energy Projects and Investment. This theme focuses on electricity generation projects, particularly solar and wind power, alongside investment flows and infrastructure expansion. Its distribution appears relatively even, suggesting stable coverage of project-level developments within the broader transition process.

Topic 4 – Air Quality and Public Exposure. Air pollution emerges as the most dominant topic in the corpus. Reporting commonly links deteriorating air quality to everyday environmental conditions and public health concerns in major urban centers, indicating a strong connection between environmental risk and lived experience.

Topic 5 – National Strategy and International Engagement. Climate and energy issues are also embedded within high-level policy narratives, including international cooperation and long-term strategic positioning. The visibility of this topic suggests that environmental discourse is not limited to local or sectoral issues but extends to national and global policy arenas.

Overall, the LDA results point to a multi-layered discourse structure in which different levels of environmental concern coexist. Immediate and observable environmental risks, particularly air pollution and local environmental degradation, receive the greatest attention, while longer-term structural themes such as energy transition and national strategy appear as secondary but stable components.

This distribution suggests that Vietnamese climate discourse is anchored in tangible and locally experienced environmental issues, which are more salient for public communication, while abstract or long-term policy narratives remain comparatively

less dominant. At the same time, the coexistence of governance, infrastructure, and international policy topics indicates that climate change is not framed as a purely environmental issue, but rather as a cross-sectoral policy domain embedded in economic development and institutional planning.

However, the relatively broad and partially overlapping nature of several topics, particularly those related to energy transition, investment, and policy, also reflects a limitation of the LDA framework. Specifically, LDA may struggle to disentangle closely related themes when it relies solely on word co-occurrence patterns.

4.2. Topic Modeling by BERTopic

The BERTopic model also yields six topics. Figure 6 presents their semantic composition using word clouds, providing a qualitative illustration of the most representative terms within each cluster.



Figure 6: Word clouds of the six topics identified by BERTopic

Topic 0 – Environmental Pollution and Air Quality. This topic accounts for the largest proportion of documents in the corpus. It centers on environmental pollution, with air quality constituting its core component. Frequent references to pollution indices, environmental conditions, and public health exposure indicate the salience of urban air concerns. The prominence of this theme is consistent with the LDA results, reinforcing its structural importance within the dataset.

Topic 1 – Renewable Energy and Power Sector Development. This cluster captures structural transformation within the electricity system. Generation capacity, grid expansion, and investment dynamics are recurrent elements, reflecting technical and infrastructural change. Compared to the corresponding LDA theme, the focus here appears more concentrated on system-level developments rather than on broader developmental narratives.

Topic 2 – Climate Change Impacts and Adaptation. This topic addresses climate-related impacts in vulnerable sectors, including agriculture, alongside adaptation measures. Its close correspondence with the adaptation theme identified by LDA suggests stability of this strand across modeling approaches.

Topic 3 – Climate Policy and International Commitments. Policy discourse emerges as a distinct cluster. Articles emphasize climate agreements, diplomatic engagement, and national commitments within international governance frameworks. In contrast to LDA, strategic positioning appears more clearly separated from implementation-related themes.

Topic 4 – Youth Engagement and Social Initiatives. This topic highlights societal participation, including youth movements and community-based initiatives. Its emergence as an independent cluster indicates that social engagement constitutes a recognizable component of climate discourse within the corpus.

Topic 5 – Provincial Energy Projects and Infrastructure. Renewable energy development at the provincial level forms a separate theme. Wind power projects and associated infrastructure are delineated from national policy discussions, suggesting that BERTopic differentiates implementation dynamics from strategic narratives more explicitly than LDA.

Across both modeling approaches, several patterns remain consistent, most notably the dominance of air pollution and environmental quality, as well as the strong presence of energy transition and climate policy themes. This convergence across models increases confidence that these topics represent structurally stable components of Vietnamese climate discourse rather than artifacts of a specific modeling approach.

4.3. Evaluation of Topic Modeling Performance

The quantitative performance of LDA and BERTopic, evaluated using coherence and diversity metrics, is reported in Table 2.

Models	LDA	BERTopic
Topic Coherence (c_v)	0.55	0.66
Topic Coherence (NPMI)	0.06	0.18
Topic Diversity	0.94	0.93

Table 2: Topic Model Performance Metrics

BERTopic achieves higher coherence scores under both measures. The c_v coherence reaches 0.66 for BERTopic, compared to 0.55 for LDA. A similar pattern is observed for NPMI coherence, where BERTopic attains a score of 0.18, whereas LDA yields 0.06. These differences indicate stronger semantic consistency among the most representative words within each topic when using the embedding-based approach.

However, the relatively low NPMI score for LDA (0.06) should be interpreted with caution, as it likely reflects a mismatch between the metric and the characteristics of the data rather than genuinely poor topic quality. NPMI relies on localized word co-occurrence and is therefore sensitive to how frequently related terms appear within a fixed window. This sensitivity makes NPMI particularly vulnerable to underestimating coherence in settings where semantic relationships are distributed across broader textual contexts rather than localized word spans.

In the present corpus, three factors systematically reduce co-occurrence signals. First, lexical synonymy fragments probability mass across multiple surface forms (e.g., *chính sách*, *quy định*, *pháp luật*), reducing pairwise co-occurrence counts even when words are semantically aligned. Second, climate discourse spans multiple subdomains, such as policy, energy, and environment, leading to domain-stratified vocabulary in which related terms rarely co-occur within the same local context. Third, corpus sparsity and high lexical diversity limit the accumulation of reliable co-occurrence statistics under standard window sizes.

These effects are further amplified by characteristics of Vietnamese text, including multi-syllabic word segmentation and lexical variability, which weaken surface-level co-occurrence signals. As a result, NPMI systematically underestimates topic coherence in this setting, particularly for probabilistic models such as LDA that rely on distributed word frequency patterns.

In contrast, the c_v metric yields substantially higher values for both models, reflecting its ability to capture broader semantic similarity beyond strict co-occurrence. The divergence between c_v and NPMI suggests that while topics may not exhibit strong local co-occurrence patterns, they remain semantically interpretable at a higher level.

Topic diversity remains high for both models, with values of 0.94 for LDA and 0.93 for BERTopic. This indicates that both approaches produce largely distinct topic representations, with minimal lexical overlap among top-ranked terms.

Taken together, the results suggest that BERTopic provides improved semantic coherence, particularly under embedding-aware evaluation, while LDA captures broader thematic structures despite lower co-occurrence-based coherence

scores. These findings highlight the importance of using multiple evaluation metrics when assessing topic models, especially in low-resource and linguistically complex settings.

4.4. Temporal Evolution of Climate Discourse

Finally, to examine the temporal dynamics of ecological discourse, we analyze the annual prevalence of LDA-derived themes over 2004–2026. Topic prevalence is computed as the yearly average of document–topic probabilities, enabling systematic identification of shifts in thematic salience over time.

Figure 7 presents the longitudinal evolution of the six themes. The results reveal distinct temporal phases: early dominance of environmental pollution narratives, followed by a rise in adaptation and governance discourse, and more recently, the expansion of energy transition and international policy framing. Several peaks align with major national and global climate-related events, suggesting that media attention responds to both structural policy developments and event-driven environmental crises.

Environmental pollution and local governance constitute a dominant theme in the mid-2000s, reaching an initial peak around 2006–2007. This period coincides with rapid industrial expansion, suggesting a possible association between increased economic activity and heightened media attention to localized environmental degradation. A renewed surge appears in 2016, aligned with the Formosa Ha Tinh marine pollution incident, followed by another increase in 2019–2020 during intensified reporting on urban air pollution in Hanoi and Ho Chi Minh City. These fluctuations indicate that pollution discourse is strongly event-sensitive, intensifying in response to acute environmental crises rather than evolving along a continuous structural trajectory.

Climate change and adaptation discourse increases markedly between 2008 and 2012, peaking around 2012. This expansion corresponds to major international policy milestones, including the IPCC Fourth Assessment cycle and the Copenhagen and Cancun negotiations, as well as domestic initiatives such as the National Target Program to Respond to Climate Change (2008) and the National Strategy on Climate Change (2011). The subsequent decline after 2013 suggests thematic differentiation, with climate concerns increasingly embedded within sectoral governance and energy-related narratives rather than remaining an isolated strand.

Energy transition follows a U-shaped trajectory. After declining until approximately 2015, coverage increases steadily from 2016 onward, aligning with

the Paris Agreement and Vietnam's Nationally Determined Contribution, suggesting that international policy developments may influence domestic media attention. A more pronounced acceleration after 2022 corresponds to COP26-related commitments and participation in the Just Energy Transition Partnership. Unlike pollution-related themes, this pattern reflects cumulative policy alignment and long-term strategic repositioning of the national energy system, rather than episodic crisis response.

Renewable energy projects and investment display sustained growth beginning around 2014, following the introduction of feed-in tariff mechanisms and renewable incentives. Stronger expansion after 2019 parallels rapid solar deployment and intensified clean energy investment. This trajectory suggests that renewable discourse is closely linked to infrastructural scaling and regulatory reinforcement, indicating structurally anchored implementation dynamics.

Air quality and public exposure exhibit a pronounced spike in 2019, coinciding with severe pollution episodes in Hanoi. The decline after 2020 likely reflects reduced mobility during the COVID-19 period, followed by stabilization as air quality becomes integrated into broader environmental governance discussions. This pattern illustrates the transition of specific environmental risks from acute public attention to normalized policy framing.

National strategy and international engagement peak around 2015 and again in 2021–2022, corresponding to the Paris Agreement process and subsequent COP26 commitments. The cyclical nature of this theme indicates temporal synchronization between national discourse and international negotiation cycles, highlighting the external anchoring of domestic climate narratives within multilateral processes.

Jointly, these patterns indicate that different thematic domains follow distinct temporal logics. Pollution-related topics exhibit short-term volatility driven by discrete events, whereas energy and policy-related themes evolve more gradually in response to institutional and regulatory processes. This divergence highlights the coexistence of reactive and structural dynamics within climate discourse.

Across these trajectories, Vietnamese climate media discourse undergoes a gradual transformation. Early coverage is characterized by reactive responses to localized environmental crises, whereas later periods show increasing institutionalization, international alignment, and structural framing of climate and energy issues. Pollution-related themes remain event-driven, while energy transition and renewable investment narratives exhibit cumulative, policy-oriented dynamics. This shift points to a transition from crisis-centered environmen-

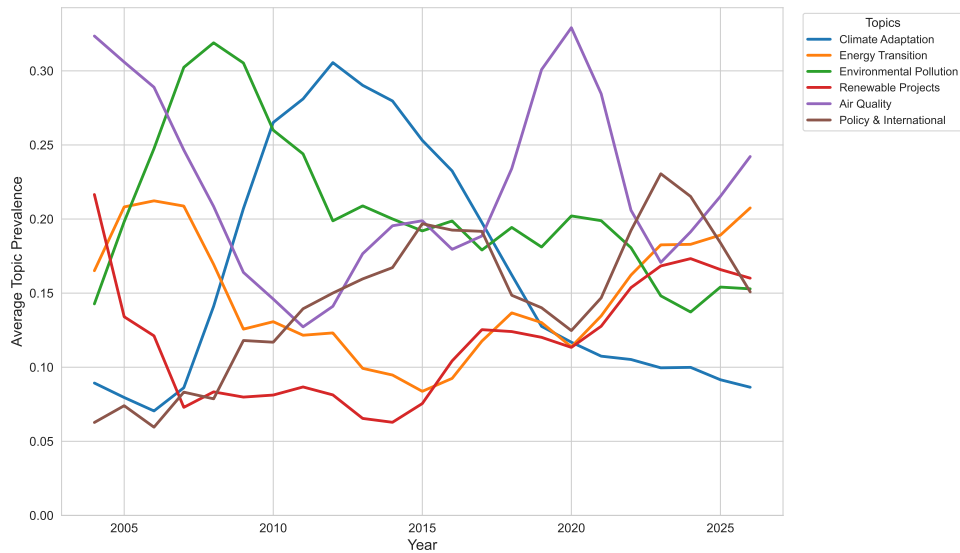


Figure 7: Temporal evolution of LDA-derived climate discourse topics in Vietnamese national media (2004–2026)

tal reporting toward a more institutionalized and policy-driven discourse, in which climate change is increasingly framed as a long-term governance and development challenge rather than a series of isolated events.

5. Conclusion

To summarise, we constructed a geographically grounded longitudinal corpus of climate-related news articles in Vietnam covering 2004–2026. The corpus was automatically collected from major national media outlets and processed using Vietnamese-specific NLP tools to ensure linguistic and geographic consistency. Using Latent Dirichlet Allocation and BERTopic, we modelled the thematic structure of the corpus and traced its evolution over time, enabling a comparison between probabilistic and embedding-based topic modeling approaches in a low-resource setting.

The results reveal stable thematic dimensions across models, including air pollution, renewable energy development, and climate governance. Longitudinal analysis highlights a gradual shift in media attention: early discourse centers on localized environmental pollution, followed by increasing emphasis on adaptation and institutional policy frameworks, and more recently, energy transition and international climate engagement. These domains exhibit distinct temporal dynamics, with pollution-related topics showing event-driven fluctuations, while energy and policy themes evolve more gradually in response to structural developments. The embedding-based approach achieves higher topic coherence while maintaining compa-

table topic diversity, indicating stronger semantic consistency.

Beyond methodological comparison, this study contributes a new Vietnamese climate corpus and provides empirical evidence on how ecological discourse evolves in a low-resource, policy-driven media environment. These findings underscore the value of combining modeling approaches to capture both high-level thematic structure and fine-grained semantic variation.

Several directions for future research emerge. The corpus could be extended to include regional outlets and social media to capture a broader spectrum of communication. Dynamic topic modeling may enable more fine-grained temporal analysis, while domain-adaptive pretraining could improve representation quality. Integrating sentiment or uncertainty modeling would further support the analysis of framing and discursive dynamics.

This study has several limitations. Vietnamese remains a comparatively low-resource language, and domain-specific pretrained models are limited. Automatic word segmentation may introduce noise affecting downstream modeling, while reliance on mainstream national media may introduce structural bias and underrepresent local perspectives. Finally, the absence of human qualitative validation limits full assessment of topic interpretability.

6. Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback and helpful suggestions. Their comments have contributed to improving the clarity and quality of this work.

7. References

- Rafiu Adekoya Badekale and Adewale Akinfaderin. 2025. [Temporal analysis of climate policy discourse: Insights from dynamic embedded topic modeling](#). *ArXiv*, abs/2507.06435.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.
- Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 2210–2216. AAAI Press.
- Rob Churchill and Lisa Singh. 2022. [The evolution of topic modeling](#). *ACM Computing Surveys*, 54(10s):1–35.
- Thi Kim Phung Dang. 2025. [Climate change communication in vietnam's online newspapers and its implications for climate actions](#). *Sustainability*, 17:1354.
- Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. [Topic modeling in embedding spaces](#). *Transactions of the Association for Computational Linguistics*, 8:439–453.
- Isabella Gagliardi and Teresa Artese. 2020. [Semantic unsupervised automatic keyphrase extraction by integrating word embedding with clustering methods](#). *Multimodal Technologies and Interaction*, 4(2):30.
- Tunahan Gokcimen and Bihter Das. 2024. [Exploring climate change discourse on social media and blogs using a topic modeling analysis](#). *Helikon*, 10(11):e32464.
- Maarten R. Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *ArXiv*, abs/2203.05794.
- F. M. Anamul Haque, Md. Abdur Rahman, and Sumon Ahmed. 2025. [Ghtm: A graph based hybrid topic modeling approach in low-resource bengali language](#). *ArXiv*, abs/2508.00605.
- Long Le and Lien-Huong Vo. 2026. [Vietnam in the climate change narratives: A discursive news values analysis of english-language news](#). *Studies in Media and Communication*, 14:267–280.
- Leland McInnes and John Healy. 2018. [Umap: Uniform manifold approximation and projection for dimension reduction](#).
- Leland McInnes, John Healy, and Steve Astels. 2017. [hdbscan: Hierarchical density based clustering](#). *The Journal of Open Source Software*, 2:205.
- Dasha Pruss, Yoshinari Fujinuma, Ashlynn R. Daughton, Michael J. Paul, Brad Arnot, Danielle Albers Szafir, and Jordan Boyd-Graber. 2019. [Zika discourse in the americas: A multilingual topic analysis of twitter](#). *PloS one*, 14(5):e0216922.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. [Exploring the space of topic coherence measures](#). WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.
- Sahil Sawant, Jinhong Yu, Kirtikumar Pandya, Chun-Kit Ngan, and Rolf Bardeli. 2022. [An enhanced bertopic framework and algorithm for improving topic coherence and diversity](#). In *Proceedings of the 2022 IEEE 24th International Conference on High Performance Computing & Communications (HPCC/DSS/SmartCity/DependSys)*, pages 2251–2257. IEEE.
- Ehsan Tashakori, Yaser Sobhanifard, Adel Aazami, and Rahim Khanizad. 2025. [Uncovering semantic patterns in sustainability research: A systematic nlp review](#). *Sustainable Development*.
- Samson Ebeneazar Uthirapathy and Domnic Sandanam. 2023. [Topic modelling and opinion analysis on climate change twitter data using lda and bert model](#). *Procedia Computer Science*, 218:908–917.
- Azmine Toushik Wasi, Wahid Faisal, Taj Ahmad, Abdur Rahman, and Mst Rafia Islam. 2024. [Dhoroni: Exploring bengali climate change and environmental views with a multi-perspective news dataset and natural language processing](#). *arXiv preprint arXiv:2410.17225*.

Greench-v1: distilling SLMs on Greenwashing Detection

Simona Scala, Federico Raspanti, Irem Demirtas,
Alessandro Pietro Bardelli, Marilena Di Bari
Michele Filannino

Prometeia spa
Piazza Trento e Trieste, 3, 40137 Bologna
{simona.scala, federico.raspanti, irem.demirtas,
alessandro.bardelli, marilena.dibari, michele.filannino}@prometeia.com

Abstract

Validating greenwashing claims in environmental, social, and governance (ESG) reports relies heavily on costly and inconsistent manual review. To address this, this paper introduces Greench-v1, a low-latency small language model (based on Qwen3-4B) that screens ESG text at the paragraph level. The model outputs a three-way classification (Greenwashing Alert, No Greenwashing, Not Relevant) paired with a concise, paragraph-grounded rationale to assist human auditors in triage and validation. The system was trained on a custom dataset of roughly 2,000 paragraphs, adapted from the ClimateBERT corpus. This dataset mitigates class imbalance through controlled paraphrasing of rare positive instances and uses GPT-4o to generate evidence-based justifications. Four training regimes were evaluated: (i) Hard distillation: Supervised fine-tuning on teacher-generated outputs. (ii) Soft distillation: Training the student to match the temperature-scaled logits of a domain-specialized Qwen3-14B teacher. (iii) Group Relative Policy Optimization (GRPO): Reward-based updates driven by exact-match alert generation. (iv) Hybrid GRPO: GRPO initialized from the hard-distilled checkpoint. Distillation and efficient policy optimization significantly improved performance over untuned baselines. Soft distillation and GRPO achieved the strongest results, increasing the "Greenwashing Alert" weighted F1-score by 36.7% and 49.0%, respectively, resulting in a deployable tool for screening large volumes of ESG narratives.

Keywords: ESG reporting, greenwashing detection, knowledge distillation, policy optimization, small language models.

1. Introduction

Over the past decade, the progressive evolution of the European sustainability regulatory framework—together with the growing demand for sustainable finance—has prompted banks, as well as other financial and non-financial institutions, to systematically integrate **Environmental, Social and Governance (ESG) factors** into their strategies, product offerings, risk management frameworks, and operating models.

The increasing relevance of sustainability thus materializes across multiple dimensions, reflecting regulatory developments, heightened stakeholder expectations, and the evolving role of financial institutions in supporting a just and orderly transition to a more sustainable economy. Among these dimensions, the safeguarding of the reputational sphere has assumed particular importance. This entails the continuous monitoring of alignment between publicly disclosed commitments (e.g. sustainability reports, climate transition plans, green bond frameworks, annual integrated reports, product-level ESG factsheets, etc.) and actual business practices, with the objective of mitigating exposure to greenwashing risk. **Greenwashing** refers to the practice whereby an institution misrepresents, exaggerates, or selectively discloses information re-

garding the environmental or sustainability-related characteristics of its products, services, or overall strategy, thereby creating a misleading perception of ESG alignment. Such practices may take the form of vague or unsubstantiated claims, incomplete or biased transparency, or the overemphasis of isolated sustainable initiatives while disregarding broader adverse environmental or social impacts. Beyond regulatory and legal implications, greenwashing poses significant reputational risks and undermines stakeholder trust and the credibility of sustainable finance frameworks.

In response to these risks, increasing attention has been devoted to the development of methodologies for greenwashing detection, particularly through the analysis of ESG-related disclosures, such as sustainability reports and other non-financial statements.

From a technological perspective, recent advancements in Natural Language Processing (NLP) have significantly enhanced detection capabilities. In particular, **Generative AI techniques, such as Large Language Models (LLMs) and Small Language Models (SLMs)** have progressively replaced traditional vocabulary-based or purely statistical approaches, enabling more nuanced, context-aware, and semantically rich analyses of textual content. These advancements have strengthened

institutions' ability to identify inconsistencies, overly generic claims, and potential misalignments between narrative disclosures and measurable sustainability performance.

2. Related Works

2.1. NLP-Based Methodologies for Evaluating ESG Disclosures

The operationalization of greenwashing detection has evolved significantly from binary assessments of truthfulness to multidimensional evaluations of **narrative strategy and communicative decoupling**.

Early NLP approaches relied heavily on simple lexicons to analyze tone and readability, finding that environmental violators often utilize more positive, verbose, yet less readable language to deflect stakeholder attention (Gorovaia and Makrominas, 2025). The field subsequently advanced with transformer-based architectures, notably **ClimateBERT**, which outperforms general-purpose models by pre-training on domain-specific climate corpora (Webersinke et al., 2021).

Building on these foundation models, scholars generally agree that detecting greenwashing requires analyzing the structural and stylistic cues of a text rather than relying solely on keyword frequencies. For instance, Binger et al. (2024) developed the "Cheap Talk Index" using ClimateBERT to identify non-specific climate commitments. To capture this complexity, researchers have proposed composite metrics like the *Greenwashing Severity Index (GSI)* and the *Green Authenticity Index (GAI)*, which combine sentiment analysis, TF-IDF weighting, and topic modeling to quantify the divergence between a firm's reported narratives and independent external evidence (Sudro and Mukhopadhyay (2025)). Despite this consensus on the utility of NLP, methodologies diverge on validation: some researchers rely purely on internal linguistic characteristics like hedging and vagueness (Livytka, 2019), whereas others argue that textual signals must be benchmarked against external performance data, such as RepRisk incidents, to conclusively prove substantive decoupling.

2.2. The advent of Large Language Models

A fundamental gap identified across the current literature is the **absence of comprehensive, gold-standard datasets** containing verified cases of greenwashing, largely due to the subjective, ambiguous, and legally sensitive nature of the phenomenon (Calamai et al., 2025). To circumvent this bottleneck, recent studies have increasingly

employed Large Language Models (LLMs) to synthesize training data or act as zero-shot evaluators. Birti et al. (2025) successfully demonstrated that augmenting manual annotations with LLM-generated synthetic data, such as controlled paraphrasing of ESG statements, significantly improves the classification accuracy of downstream models.

However, relying on massive, general-purpose LLMs introduces profound vulnerabilities into the auditing pipeline. Chuang et al. (2025) empirically demonstrated that LLMs can be weaponized by corporations to seamlessly generate highly convincing greenwashed responses that evade standard detection mechanisms, unless strict accuracy constraints are externally imposed. These vulnerabilities highlight a critical gap in the literature: massive, **black-box LLMs are too resource-intensive, unpredictable, and easily manipulated** to serve as reliable, large-scale financial auditing tools.

2.3. The Shift Toward more efficient and explainable Models

To achieve the deep reasoning capabilities of massive LLMs without their computational overhead and unpredictability, the literature is pivoting toward **Small Language Models (SLMs)** enhanced via **Knowledge Distillation (KD)** and reinforcement learning to democratize language evaluation. KD addresses the limitations of standard training by transferring the continuous probability distributions (logits) of a massive teacher model into a compact student model, allowing the SLM to internalize complex domain nuances while remaining accessible to researchers with limited compute resources. Recent advancements emphasize **"Distilling Step-by-Step,"** wherein student models are trained on natural language chain-of-thought rationales generated by the teacher, enabling them to outperform much larger models while utilizing fewer parameters (Hsieh et al., 2023). Furthermore, post-training alignment techniques, such as **Group Relative Policy Optimization (GRPO)**, allow these models to learn from verifiable, multi-objective rewards, ensuring that their evaluation outputs remain factually grounded, structurally sound, and scalable for reproducible research without the prohibitive costs of traditional reinforcement learning.

3. Greench-v1

The aforementioned research issues (the necessity for multi-dimensional linguistic evaluation, the critical bottlenecks of data scarcity and LLM vulnerabilities, and the promise of KD-optimized SLMs) directly frame the specific problem addressed by **Greench-v1**. While current approaches either rely on simplistic, static proxies or brittle, computation-

ally expensive LLMs that lack the transparency required for rigorous financial assurance, *Greench-v1* addresses these limitations through a compact, locally deployable architecture that enables low-latency inference, full training transparency, and structured label-and-rationale outputs suited for human-in-the-loop ESG auditing pipelines.

Processing Pipeline

The deployed model utilized a custom soft-distillation trainer. Additionally, the prompt structure was optimized by explicitly defining greenwashing in the system prompt. The compiled *Greench-v1* model is operationalized via a web-based interface that strictly orchestrates document processing from ingestion to output generation. The procedural workflow is as follows:

1. **Input Ingestion:** A user uploads a target sustainability document (in PDF format) into the system interface.
2. **Execution and Chunking:** Upon user initiation, the system performs Optical Character Recognition (OCR) and segments the entire document into discrete paragraphs.
3. **Iterative Processing:** The core algorithm analyzes the text sequentially, outputting the results paragraph by paragraph to the interface.
4. **Classification Formulation:** For every processed paragraph, the model computes a classification and corresponding justification. It assigns "Not relevant" to text lacking environmental claims, "No greenwashing" to substantiated claims (e.g., factual references to operational certifications), and triggers a "Greenwashing alert" when identifying broad, unmeasured environmental commitments lacking concrete action (Fig. 1)
5. **Synthesis and Export:** Once paragraph-level iteration concludes, the system aggregates a summary of the analysis at the bottom of the interface. Finally, the granular results, including text chunks, labels, and justifications, can be exported into standard, machine-readable file formats (such as CSV or JSON) for external auditing or reproducibility (Fig 2).

4. Data Collection and Preprocessing

4.1. Primary Dataset Construction

The dataset construction commenced with the `climatebert/climate_specificity` dataset, a binary classification task determining whether a given climate-related paragraph is specific or

non-specific in nature (Webersinke et al., 2021), yielding 1,320 paragraphs. These instances were mapped to formulate an initial dataset categorized by a binary classification schema into "POSSIBLE GREENWASHING" and "NO GREENWASHING" labels. To incorporate a null class, 660 paragraphs (one third of the total) explicitly classified as non-environmental claims were sampled from the `climatebert/environmental_claims` dataset, which provides a binary classification task identifying whether a given sentence constitutes an environmental claim or not, and designated as "NOT RELEVANT" (Webersinke et al., 2021).

To mitigate class imbalance, the underrepresented "POSSIBLE GREENWASHING" class was augmented via controlled paraphrasing. Specifically, each positive instance was paraphrased three times using GPT 4.1, to expand the minority class distribution and match the distribution of the null and negative classes. Subsequently, evidence-based textual justifications for each sample were synthesized utilizing GPT-4.1, which generated a rationale strictly conditioned on the source paragraph and its corresponding label. The final pre-processing pipeline yielded a curated corpus of approximately 2,000 paragraphs, wherein each entry comprises the source text, a discrete three-way classification label, and an accompanying natural language justification.

The resulting corpus was partitioned into training and evaluation sets via a stratified 80/20 split, yielding approximately 1,600 paragraphs for training and 400 for evaluation, with class proportions preserved across both subsets.

4.2. Supplementary ESG Corpus for Hard Distillation

In addition to this primary corpus, a supplementary dataset was constructed exclusively for the Hard Distillation + GRPO training paradigm described in Section 5.1. This dataset was assembled from 30 real-world ESG reports, segmented at the paragraph level. Each paragraph was annotated by GPT-4.1 according to the same three-way classification schema ("POSSIBLE GREENWASHING", "NO GREENWASHING", "NOT RELEVANT"), accompanied by a corresponding natural language justification. The objective of this corpus is to expose the model to a broad distribution of authentic ESG discourse prior to policy optimization, thereby providing domain familiarization at scale before the reward-driven fine-tuning stage.

Field	Content
Source Text	<i>In 2022, Mango joined the network of the Ellen MacArthur Foundation, an international charity dedicated to promoting the circular economy. This vision aims to eliminate waste and pollution, encourage the circulation of products and materials, and regenerate nature, with the goal of transforming the economic system towards a more sustainable one.</i>
Label	POSSIBLE GREENWASHING
Justification	This paragraph presents a general statement about Mango’s association with the Ellen MacArthur Foundation. However, there is no specific mention of concrete actions taken nor detailed information regarding how these objectives have been implemented or achieved. The text relies on broad, unsubstantiated references to organisational commitments rather than tangible outcomes.

Figure 1: Example paragraph-level output from Greench-v1 on a Mango ESG report, flagged as POSSIBLE GREENWASHING due to the lack of concrete, measurable environmental commitments.

Category	Count
Total paragraphs	29
Greenwashing alerts	6
Not relevant	11
Clean paragraphs	12

Figure 2: Example summary statistics produced by Greench-v1 after processing a Mango ESG report. Results can be exported in CSV or JSON format for external auditing or reproducibility.

5. Methodology

5.1. Experiments

All training and inference experiments were conducted on a single instance (24 vCPUs, 220 GB RAM, 64 GB disk, equipped with one NVIDIA A100 40 GB GPU). Inference on the held-out evaluation set of approximately 400 paragraphs completed in under 5 minutes.

To rigorously select the optimal learning strategy for the Greench-v1 architecture, four distinct training paradigms were investigated (Fig.3).

First, a **hard distillation** approach was implemented, wherein the Qwen3-4B model was supervised directly on the ground-truth targets, encompassing both the discrete classification labels and their associated textual justifications.

Second, a **soft distillation** methodology was examined. This paradigm involved the initial fine-tuning of a Qwen3-14B teacher model, followed by the optimization of the Qwen3-4B student using a composite loss function. This function integrates the cross-entropy loss over the gold tokens with the **forward Kullback-Leibler divergence** computed between the temperature-scaled logits of the teacher and the student. The formulation incor-

porates hyper-parameters α and T to modulate the relative weighting of the teacher’s distribution against the ground-truth answers, and to adjust the distributional entropy of the teacher’s signals. However, implementing this approach incurred a soft-distillation adaptation cost, as the custom script required time to be adapted from vision models to transformers.

Third, **Group Relative Policy Optimization (GRPO)** was evaluated using a discrete reward mechanism. The reward function strictly assigned +1 for an exact syntactic match with the ground-truth label and +0 otherwise. Despite this straightforward assignment, managing GRPO complexity was challenging, as defining an appropriate and stable reward function proved difficult. Throughout the GRPO training phase, the model decoded 10 distinct candidate sequences per optimization step, with the policy gradient algorithm updating the network parameters to maximize the averaged expected reward across the generated group.

Finally, a hybrid **Hard Distillation + GRPO** paradigm was investigated to assess whether domain familiarization prior to policy optimization could further improve classification performance. In this regime, Qwen3-4B was first subjected to hard distillation on the supplementary ESG report corpus described in Section 4.2, exposing the model to a large and diverse distribution of authentic ESG narratives. The resulting checkpoint was subsequently used to initialize GRPO training on the primary dataset, applying the same exact-match reward function as in the standalone GRPO condition. The rationale underlying this two-stage procedure is that broad domain exposure during the distillation phase may yield a more favorable **parameter initialization** for reward-driven fine-tuning, potentially accelerating convergence and improving robustness.

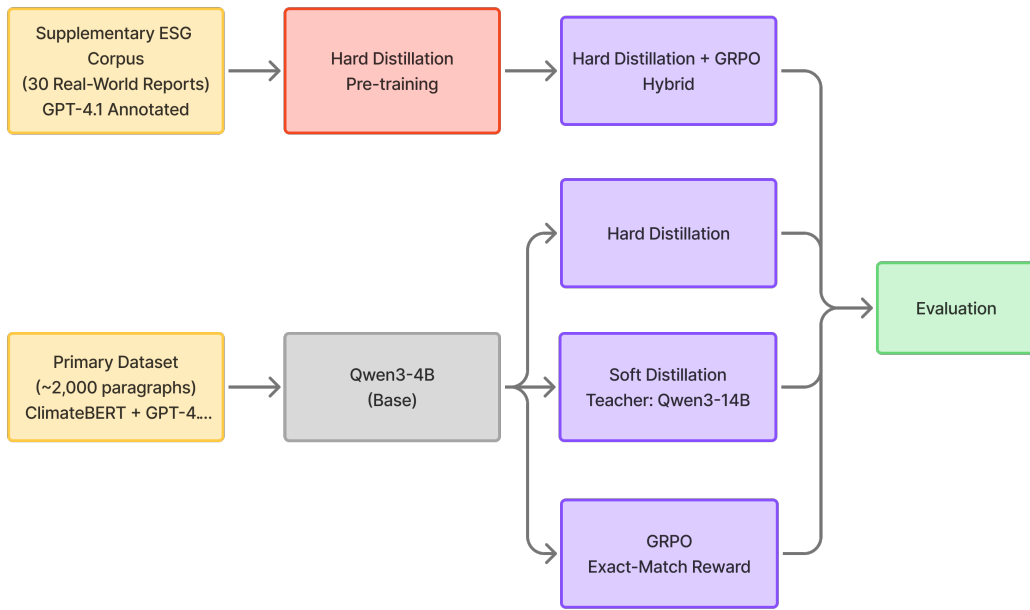


Figure 3: Overview of the four training paradigms evaluated for Greench-v1: Hard Distillation, Soft Distillation (with Qwen3-14B as teacher), GRPO with exact-match reward, and a Hybrid Hard Distillation + GRPO approach initialized from the supplementary ESG corpus checkpoint.

6. Results

6.1. Evaluation Metrics

To assess model performance across both output dimensions, two complementary metrics were employed. For the discrete classification task, the **weighted F1-score** was adopted to evaluate Greenwashing Alert predictions.

For the free-text justification output, **BERTScore** (Zhang et al., 2020) was used to measure the semantic similarity between the model-generated rationale and the reference justification. BERTScore leverages contextual embeddings from a pre-trained language model to compute token-level similarity via greedy matching, producing precision, recall, and F1 estimates that capture meaning beyond **surface-level lexical overlap**. The F1 variant of BERTScore is reported throughout.

For the discrete classification task, the **weighted F1-score** was adopted to evaluate Greenwashing Alert predictions. Although the evaluation set is approximately balanced across classes, weighted F1 was retained for consistency with early experimental runs conducted on the original imbalanced corpus; under balanced conditions, the weighted and macro-averaged F1-scores are functionally equivalent.

6.2. Quantitative Evaluation

Table 1 reports the Greenwashing Alert weighted F1-score and the Justification F1-score for all evaluated baselines and distilled model variants.

The results reveal a consistent and substantial **performance gap** between the untrained baselines and the distilled model variants across both evaluation dimensions. Among the baselines, the untuned Qwen3-4B achieves the highest Greenwashing Alert weighted F1-score of 0.49, which is notably counter-intuitive given that Qwen3-32B—a considerably larger model—underperforms at 0.45. This suggests that **raw parameter count does not directly confer advantage** in this domain-specific task, and that a smaller model already fine-tuned for instruction-following may be more amenable to zero-shot greenwashing classification. This suggests that **raw parameter count does not directly confer advantage** in this domain-specific task; the marginal gap of 0.04 between the two models is small enough to be attributable to stochastic variation in decoding rather than a systematic capability difference, and both models should be regarded as equivalent zero-shot baselines. `claude-4.5-sonnet` records the weakest classification score (0.34) despite matching Qwen3-4B on the Justification F1 metric (0.62), indicating a dissociation between its capacity to generate coherent ratio-

Model	Greenwashing Alert (Weighted F1)	Justification (F1)
<i>Baselines</i>		
Qwen3-4B	0.49	0.62
Qwen3-32B	0.45	0.58
claude-4.5-sonnet	0.34	0.62
<i>Distilled Models</i>		
Qwen3-4B-hard-distillation	0.54	0.59
Qwen3-4B-soft-distillation	0.67	0.64
Qwen3-4B-grpo	0.73	0.62
Qwen3-4B-hard-distillation + grpo	0.60	0.62

Table 1: Performance of baseline and distilled models on the Greenwashing Alert weighted F1-score and Justification F1-score. Bold entries indicate the best result per column.

nales and its ability to reliably assign the correct alert label under zero-shot conditions. Indeed, we find that `claude-4.5-sonnet` was too eager at classifying paragraphs as potential greenwashing.

Across the distilled variants, all four training paradigms improve upon the best baseline on the Greenwashing Alert metric, confirming that **task-specific supervision** is a necessary condition for reliable greenwashing triage. Hard distillation yields a modest improvement to 0.54, demonstrating that supervised fine-tuning on teacher-generated gold labels alone is insufficient to fully capture the distributional nuances of the classification task. The most pronounced gains are attributable to soft distillation and GRPO. Soft distillation achieves a weighted F1-score of 0.67—a relative improvement of 36.7% over the best baseline—and simultaneously records the highest Justification F1-score of 0.64, suggesting that exposure to the teacher’s full **token-level probability distribution** yields richer, more transferable representations that benefit both classification accuracy and rationale quality. GRPO attains the strongest Greenwashing Alert score of 0.73, corresponding to a 49.0% relative gain over the best baseline, underscoring the effectiveness of reward-driven policy optimization in sharpening the model’s sensitivity to the exact syntactic structure of alert labels. However, GRPO does not improve the Justification F1-score beyond 0.62, implying that a **binary exact-match reward** is sufficient to steer classification behavior but does not incentivize qualitative improvements in the accompanying rationale.

The Hard Distillation + GRPO hybrid achieves a Greenwashing Alert weighted F1-score of 0.60 and a Justification F1-score of 0.62. While this represents a meaningful improvement over the hard distillation baseline (0.54), it falls short of the standalone GRPO result (0.73). This outcome suggests that initializing from a checkpoint trained on the large ESG corpus does not provide a more favorable starting point for policy optimization than the default instruction-tuned initialization. A plausible explanation is that hard distillation on GPT-4.1-

annotated ESG reports, while broadening domain coverage, may simultaneously introduce **labeling noise** or stylistic biases that partially interfere with the reward signal during the subsequent GRPO phase. Nonetheless, the hybrid approach does match the standalone GRPO on justification quality (0.62), indicating that domain pre-exposure does not degrade rationale generation.

6.3. Qualitative Evaluation

To complement the quantitative evaluation, we examine representative predictions to characterize the system’s behaviour across correct and incorrect classifications.

True Negative. The model correctly assigns No GREENWASHING DETECTED to a paragraph discussing climate-related physical risks, such as the expected southward extension of cyclones in Australia and their implications for asset pricing. The justification correctly identifies that the text offers factual, risk-oriented disclosure without any attempt to portray the institution as environmentally responsible, demonstrating that the model has learned to distinguish risk reporting from sustainability marketing.

True Positive. The model correctly flags a paragraph in which an investor states a general intention to contribute to a circular economy, citing their network, scale, and influence, without providing any measurable targets, concrete actions, or verified outcomes. The generated justification accurately identifies the reliance on aspirational language and the absence of substantiated commitments as the key markers of greenwashing.

False Positive (Type I Error). The model incorrectly raises a GREENWASHING ALERT for a paragraph describing a \$25 million investment in electric vehicle charging infrastructure, including specific pilots in Minnesota and plans for fleet conversion.

While the target label is No GREENWASHING DETECTED, the model penalizes the use of forward-looking language (“expect to expand”) and the absence of verified outcome metrics. This suggests a systematic tendency to over-flag paragraphs that mix concrete actions with future-oriented commitments.

False Positive (Type I Error). A second false positive arises on a paragraph reporting dedicated ESG headcount across Finance, Communications, and Risk departments, including ten staff working on climate-related risk methodologies. Despite the specificity of the staffing figures, the model flags the paragraph due to the lack of measurable outcomes or verified results linked to these activities. This indicates that the model may insufficiently reward operational transparency in the absence of explicit performance metrics.

7. Conclusions

This paper presented Greench-v1, a lightweight, proprietary small language model for paragraph-level greenwashing detection in ESG disclosures. Empirical results demonstrate that task-specific supervision via knowledge distillation and policy optimization substantially outperforms zero-shot baselines, with soft distillation and GRPO yielding relative improvements of 36.7% and 49.0% on the Greenwashing Alert weighted F1-score, respectively.

Several properties of Greench-v1 make it particularly well-suited for deployment in real-world ESG auditing pipelines. First, its compact 4B-parameter architecture enables low-latency inference at scale, rendering it feasible for organizations lacking access to high-performance compute infrastructure. Second, as a proprietary, locally deployable model, Greench-v1 is not subject to undisclosed behavioral updates or capability drift associated with commercial API-based systems, ensuring reproducibility and auditability over time. Third, full transparency over the training corpus, derived from the ClimateBERT dataset with controlled augmentation, permits principled assessment of the model’s domain coverage and potential biases, a prerequisite for regulatory-grade financial assurance. Finally, the modular paragraph-level interface, with its structured label-and-rationale output schema, serves a dual purpose. As a pre-publication writing assistant, it enables ESG report editors to iteratively refine paragraph wording, reduce exposure to reputational and regulatory risk, and ensure claims are substantiated prior to disclosure. As a post-publication triage component, it integrates within broader, multi-stage ESG analysis systems, including document-level aggregation pipelines or human-

in-the-loop auditing workflows.

From a broader financial perspective, Greench-v1 represents a methodologically novel contribution to the intersection of NLP and financial compliance. Unlike general-purpose LLMs deployed as black-box evaluators, Greench-v1 introduces a transparent, auditable, and resource-efficient framework for automated narrative scrutiny that directly addresses the operational and regulatory constraints faced by financial institutions. The combination of knowledge distillation and policy optimization offers a replicable methodology for developing domain-specialized compliance tools beyond greenwashing detection. It is applicable, for instance, to the screening of prospectuses, loan documentation, or product-level ESG factsheets for regulatory adherence. In an environment of increasingly stringent disclosure requirements, the ability to deploy locally auditable, low-latency models at scale constitutes a meaningful advance over purely manual or API-dependent review processes. Greench-v1 thus demonstrates that compact, distilled language models can serve as credible first-line screening tools within institutional risk management and compliance workflows, reducing both the cost and latency of ESG-related due diligence.

Qualitative analysis reveals that the primary failure mode of the best-performing model consists of false positives arising from paragraphs that combine concrete actions with forward-looking language or report operational transparency without explicit performance metrics. Addressing this limitation is a key priority for future development.

Future work should investigate the extension of the reward function in GRPO to encompass rationale quality metrics, as well as the evaluation of Greench-v1 on external, independently curated greenwashing benchmarks to assess generalization beyond the ClimateBERT domain. Additionally, extending the framework to multimodal greenwashing detection represents a promising avenue, as ESG documents frequently embed visual elements—such as images, graphs, and charts—that may convey or obscure sustainability claims independently of the accompanying text; incorporating such modalities could yield a more comprehensive assessment of disclosure integrity.

8. Acknowledgements

The authors would like to thank our colleagues Lorenzo Proserpi and Michele Cimino for their valuable insights and constructive feedback on the overall framework.

9. Bibliographical References

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. [How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk](#). *Journal of Banking Finance*, 164:107191.

M Birti, F Osborne, and A Maurino. 2025. Optimizing large language models for esg activity detection in financial texts. arxiv. *arXiv preprint arXiv:2502.21112*.

Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M Suchanek. 2025. Corporate greenwashing detection in text—a survey. *arXiv e-prints*, pages arXiv–2502.

Marianne Chuang, Gabriel Chuang, Cheryl Chuang, and John Chuang. 2025. [Judging it, washing it: Scoring and greenwashing corporate climate disclosures using large language models](#). In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 17–31, Vienna, Austria. Association for Computational Linguistics.

Nina Gorovaia and Michalis Makrominas. 2025. [Identifying greenwashing in corporate-social responsibility reports using natural-language processing](#). *European Financial Management*, 31(1):427–462.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Inna Livytska. 2019. The use of hedging in research articles on applied linguistics. *Journal of language and cultural education*, 7(1):35–53.

Protima Nomo Sudro and Shreya Mukhopadhyay. 2025. Greenwashing detection with causal explanation: A novel multi-layered approach. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

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

10. Language Resource References

Webersinke, Nicolas and Kraus, Mathias and Bingler, Julia Anna and Leippold, Markus. 2021.

Analyzing Environmental Discourse through Construction-Based Pattern Extraction

Elisa Chierchiello¹, Eliana Di Palma¹, Ludovica Pannitto², Cristina Bosco¹

¹ University of Torino

² University of Bologna

{elisa.chierchiello, eliana.dipalma, cristina.bosco}@unito.it, ludovica.pannitto@unibo.it

Abstract

Environmental issues are at the centre of a debate currently taking place across all communication channels. This paper provides an analysis of texts in which these issues are discussed, with the novelty of applying a methodology that enables the extraction and comparison of different narratives and points of view. The texts used in this study are the English *Living Planet Reports* published biennially by the WWF from 2014 to 2024. The methodology is based on the extraction of constructions – patterns collected in the English construction CASA – which allow us to identify differences in the presentation of the issues discussed in the analysed texts. Our results show that this methodology can be very helpful in the comparative analysis of texts to reveal different perspectives, for example, to observe diachronic variations.

Keywords: environmental discourse, construction grammar, Grew, Universal Dependencies, corpus linguistics

1. Introduction

Environmental issues are increasingly debated by various public and private entities, as well as citizens, in all countries and across multiple communication channels and media. The complexity inherent to most of these issues, such as pollution, loss of biodiversity, and global warming, makes especially hard to discuss them or to recognize and compare the diverse opinions expressed when they are discussed. Different perspectives can express opinions related to varying urgencies and perceptions, with different impact on the reader and listener. For this motivation, not only the life sciences, but also disciplines such as computational linguistics are expected to address the challenges posed by environmental discourse by providing new approaches for their analysis.

In recent years, constructionist approaches have been increasingly applied across a wide range of subfields. In particular, Construction Grammar (CxG, Fillmore et al. 1988; Goldberg 2013) frameworks have proven effective in the study of information structure and discourse patterning (e.g., Lambrecht 1994; Östman 2005), as well as in research on context and pragmatics (e.g., Jing-Schmidt 2017). They have also been successfully employed in dialogue and conversation analysis (e.g., Linell 2009), and in investigations of register and genre (e.g., Fischer and Aarestrup 2021; Antonopoulou and Nikiforidou 2011; Nikiforidou and Fischer 2015).

A key principle of CxG is that every human language can be understood as a richly structured, semantically motivated, and constantly evolving inventory of symbolic units known as *constructions*. Each construction consists of a pairing between a

specific form and a corresponding meaning. These form–meaning pairings vary widely in both size and level of abstraction: they range from simple morphemes such as *book* to larger phrasal units like *you can't teach an old dog new tricks*, and from fully fixed expressions such as *break the ice* to highly schematic patterns like $SUBJ_{agent} V OBJ_{recipient} OBJ_{theme}$ (i.e., the English ditransitive construction, as in *She gave him a gift*).

In this paper, we apply a novel methodology for the automatic extraction of constructions to texts discussing environmental issues. The central hypothesis is that constructions can provide valuable insights into how environmental topics are framed. As a preliminary step, the study focuses on developing this extraction method, which is innovative in both environmental linguistics and the computational study of construction grammar. It then presents an analysis of constructions identified in the WWF corpus, offering initial evidence of the method's potential for examining environmental narratives and discourse.

As far as the texts, we use the collection of WWF's reports called *Living Planet Reports* (WWF-LPR) which are published every two years on the website of the international organization. As a case study we use the English data published from 2014 to 2024. These reports have been used in previous studies (Bosco et al., 2025; Chierchiello et al., 2025) and made available in Universal Dependencies (UD, De Marneffe et al. 2021) format in the multilingual treebank for environment TREEn (Pagano et al., 2025). UD is a cross-linguistically consistent framework for morpho-syntactic annotation that represents sentences as dependency trees, providing standardized information about grammatical rela-

tions and lexical features.

Taking advantage of the available morphological and syntactic annotations, our approach is grounded in the Construction Grammar framework. Considering that the environmental discourse is rich in recurrent constructions that carry stance, causality, and evaluation, we developed scripts to extract all the available constructions from the collected texts benefiting from their availability in UD format. Nevertheless, constructions (Lyngfelt et al., 2018) are not directly searchable in UD corpora due to the lack of a shared UD-level formalization. To address this limitation, we focus on the English CASA (Herbst and Hoffmann, 2018) constructions inventory and we operationalize the defined constructions as *Grew* graph patterns (Guillaume, 2019) over UD parses.

The main **contributions** of this study are:

- A reproducible method to operationalize CASA constructions for the analysis of discourse;
- A validated constructional inventory that makes it possible to systematically study how sustainability is expressed in language;
- A corpus-based analysis of WWF’s reports illustrating how constructional patterns shape environmental narratives.

2. Related Work

2.1. Computational Construction Grammar

A key theoretical aspect of CxG is the idea of language as a network of constructions: a *constructicon*. In recent years, however, another type of constructicon has been coming to the fore: an archive of construction descriptions, a sort of construction dictionary, halfway between CxG and lexicography (Lyngfelt et al., 2018). The idea was first introduced by Fillmore (2008), who initiated the development of an English constructicon (Fillmore et al., 2012; Lee-Goldman and Petruck, 2018) as a complement to the English FrameNet. Since this first step for English (followed by more recent resources such as *The English Constructicon* Perek and Patten 2019 and CASA, employed in this work, Herbst and Hoffmann 2018), much work has been done in the creation of constructicons for different languages: Brazilian Portuguese (Laviola et al., 2017), Swedish (Gruzitis et al., 2015), German (Ziem and Boas, 2017), Russian (Janda et al., 2018), and Japanese (Ohara, 2013), and more currently being developed.

Constructicons are systematic collections of language-specific constructions, in which the syntactic form and the corresponding meaning are

reported for each construction. Unfortunately different notations and labels were used in different constructions that make difficult to generalize constructions both intralinguistically and interlinguistically. Another challenge is the automatic extraction of constructions from texts. Similarly to our approach, (Lorenzi et al., 2023) conducted experiments on datasets annotated with UD, aiming to understand the impact of UD and frame information on the labeling of constructions elements, focusing on clausal constructions in Brazilian Portuguese (Torrent et al., 2018). Using both LSTM and Transformer-based models, they show that UD annotations and frame information can boost performance. However, their work analyzed only clausal constructions, whereas our methodology refers to all types of constructions, enabling us to base discourse analysis on all the extracted constructions.

2.2. *Grew*

*Grew*¹ (Guillaume, 2019) is a graph rewriting tool designed to manipulate and query linguistic representations encoded as graph structures. Through the *Grew* query language, users can search for occurrences of specific graph patterns within treebanks by specifying constraints on the nodes and relations involved in the query. Unlike simple keyword search, *Grew* supports graph-level queries, allowing users to precisely describe the linguistic structures they wish to retrieve from a corpus. Both Weissweiler et al. 2024 and Pisciotto et al. 2025; Pannitto et al. 2024 introduced *Grew* queries as possible formalizations of construction-like objects.

2.3. NLP for Environmental Discourse

The application of NLP to environmental discourse has been mainly centered on issues related to climate change, while Sentiment Analysis has been the most used technique. According to a systematic review (Ibrohim et al., 2023) studies that have dealt with environmental issues using NLP are few and based on very coarse-grained polarity analysis, even for English. In a similar vein, Stede and Patz 2021 examine the application of Sentiment Analysis to climate change discourse, pointing out that different groups (the general public, policymakers, and scientists) use different genres, registers, and terminologies. Other studies applied Sentiment Analysis to topics such as smart cities, urban policies, and energy debates on social media platforms (Du et al., 2020).

Environmental NLP has steadily emerged as a research field, with workshops, shared tasks, and

¹<https://grew.fr/>

domain-related language models such as ClimateBERT (Webersinke et al., 2022). However, model-driven and topic-focused approaches are used without much emphasis on the systematic study of linguistic constructions. Released corpora include an Italian resource for environmental analysis (Grasso et al., 2024a), some multilingual efforts (Bosco et al., 2023), and diachronic climate-focused corpora for English (Grasso et al., 2024b). Among the very few annotated resources the InsightsNet Climate Change Corpus (ICCC, Bartsch et al. 2023) has to be mentioned, which provides metadata, morphological, syntactic, and named entities, all automatically annotated. The only reported treebank about environmental issues is currently TREEN (Pagano et al., 2025) which is released in UD format.

Building on a corpus derived from WWF's reports, which spans a broader range of environmental issues instead of simplifying discourse analysis to sentiment polarity classification, we take a constructionist approach to identify recurring patterns of linguistic expressions that constitute environmental discourse.

3. Data: WWF's *Living Planet Reports*

The corpus used for our research is based on the WWF-LPR², a set of policy reports on biodiversity loss, ecosystem degradation, climate change, resource management, and sustainability transition that have been widely distributed across the globe in a large number of languages also mirroring cultural differences. The WWF-LPR presents an environmental story that encompasses ecological, economic, and socio-political elements, that makes them a strategic tool in the shaping of global environmental communication. As mentioned above, this corpus is part of TREEN (Pagano et al., 2025), a multilingual treebank project focused on environmental discourse and annotated within the UD framework. However, in the current contribution, we exclusively focus on the English subcorpus of this resource, and we take into account only data generated in the time span 2014-2024. The resulting dataset comprises six report editions, totaling 64,345 tokens and 2,713 sentences. Texts are automatically annotated and manually checked to conform to UD guidelines, which allows fine-grained morpho-syntactic querying. This layer of structured annotation enables us to query constructional patterns via *Grew* queries and examine how environmental narratives are linguistically encoded. Through the lens of a UD-annotated environmental corpus, our goal is to look beyond the surface level of sentiment or stance classification and instead

²<https://livingplanet.panda.org/>

explore the structural patterns by which environmental meaning and argumentation are realized.

4. Method: From CASA to UD-based *Grew* Queries

In this study, the term “construction” does not refer to an abstract syntactic unit in the traditional grammatical sense, but to a formally operationalized pattern derived from CASA and implemented as a UD-based graph query in *Grew*. This section outlines how qualitative constructional descriptions were mapped onto dependency-based graph patterns.

4.1. Selection of Constructions

The constructions were taken directly from the CASA Constructicon³ (Herbst and Hoffmann, 2018), which was chosen because of its wide coverage and availability as a structured inventory in a machine-readable format. Moreover, the qualitative descriptions of constructions can be easily transformed into dependency-based patterns. The inventory used is the revised version from December 2025. We mapped each construction's slot onto UD relations and pertinent morphosyntactic properties after looking at its syntactic schema and semantic function. As a result, constructional descriptions and graph-based representations were explicitly aligned. The result is a CASA inventory that has been UD-operationalized, allowing constructs to be automatically extracted using graph rewriting techniques.

4.2. Operationalization Principles

The operationalization of the CASA constructions into UD-based graph patterns followed a set of consistent principles.

First, the syntactic head of a construction corresponds to the UD head of a phrase, since we adopted the same UD head-selection criteria.

Second, whenever possible constructional slots were mapped to dependency relations. Other morphosyntactic constraints (e.g., *VerbForm=Inf*, *Degree=Cmp*, clause type) were encoded as feature-based restrictions, and argument roles like *subject*, *object*, *complement*, or *modifier* were translated into their corresponding UD relations. Also, we gave structural arrangements precedence over linear order and instead of capturing surface sequences, the patterns record hierarchical relationships since UD encodes syntactic structure regardless of word order.

Lexical restrictions, lastly, were only incorporated

³<https://constructicon.de/>

when they were specifically mentioned in the original composition (e.g., fixed coordinators like *and*, certain auxiliary verbs). Discourse-pragmatic subtleties and intonation, which are not technically stored in UD, were not operationalized.

4.3. Query Design in Grew

Core relations. The queries were written applying a limited but well chosen set of core UD dependent relations that emerge across constructions. We relied on relations like `root`, `subj`, `obj`, `iobj`, `xcomp`, `ccomp`, `advcl`, `acl`, `nmod`, `obl`, `amod`, `compound`, `conj`, `cc`, `mark`, and `case`. The main idea was to give each construction a structure that is easy to recognise, regardless of word order or language.

Constraints. Structural relations were merged with morphosyntactic restrictions expressed over UD features, such as `upos`, `lemma`, `VerbForm`, `Tense`, `Degree`, and `PronType`. The construction's grammatical profile is refined by feature constraints, whereas dependency relations capture its basic structure. These restrictions enable us to distinguish, for instance, between declarative and interrogative sentence types, comparative forms and positive adjectives, and infinitive and finite clause complements.

Coverage vs precision trade-offs. Designing the queries required a constant balance between recall and precision. Optional elements (e.g., modifiers), attachment ambiguities, and cross-linguistic variation can easily lead to under-generation (missing valid instances) or over-generation (including structurally similar but constructionally unrelated cases). To solve this, we gave priority to the fundamental structural arrangement that characterizes every structure, treating auxiliary components as optional whenever feasible. Especially in big, diverse corpora, this approach tolerates syntactic heterogeneity while maintaining the constructional backbone.

Figure 1 illustrates the operationalization of construction [66], *indefinite_plural*. In UD terms, the construction is identified as a plural noun that does not license a determiner dependency. The absence of a `det` relation therefore serves as the formal constraint distinguishing indefinite plural noun phrases from definite or determiner-marked ones.

Indefinite Plural NP Construction

NOUN Number = Plur

Forbidden dependency:

$\nexists D, \text{det} (\text{NOUN}, D)$

Figure 1: Indefinite Plural NP Construction, CASA ID: 66

The corresponding Grew query implementing this constraint is shown below:

```
pattern {
  N [upos="NOUN", Number="Plur"];
}
without {
  D [];
  N -[det]-> D;
}
```

The query, performed on a set of UD parsed sentences, returns a list of hits each composed by the sentence ID and a list of labeled nodes matching the ones declared in the Grew pattern.

5. Construction Inventory and Documentation

In CASA, each construction is identified by an ID and the corresponding documentation can be accessed via the CASA Constructicon website.⁴ In the remainder of this paper, CASA IDs will be reported in square brackets.

5.1. Construction Families Covered

The construction inventory examined in this study includes recurrent syntactic patterns relevant to the encoding of environmental narratives in policy discourse. We selected construction families covering aggregation, comparison, modification, coordination, and argument structure, which are central to the rhetorical strategies of this genre.

In particular, we examine 133 constructions, among:

- **Noun Phrase constructions**, like pre- and post-modification structures, which play an important role in the compact formulation of environmental ideas (e.g., "*environmentally friendly*", "*causes of unsustainable development*").

⁴[https://constructicon.de/constructions/\[ID\]](https://constructicon.de/constructions/[ID]), where [ID] is the construction identifier.

- **Connection constructions**, like [146] additive, [147] alternative, and [145] asyndetic coordination patterns, often employed to group drivers, impacts, and policy responses (e.g., *"positive, equitable, and sustainable"*).
- **Comparative constructions**, including [116] morphological comparatives and also *more/less ... than* patterns, encoding escalation, urgency, and evaluation (e.g., *"higher than it would be"*).
- **Argument-structure and causative patterns**, which are part of the representation of agency and responsibility in environmental processes (e.g., *"world must move towards a just transition"*).

5.2. Release Format

The release consists of the outputs of the `Grew` queries applied to the WWF-LPR corpus.

The outputs are provided as structured tables (in spreadsheet format), where each row corresponds to a matched instance and includes information on the construction label, sentence identifier, matched token span, head token, and relevant slot assignments.

File naming follows the original CASA construction identifiers, ensuring traceability between the construction inventory and the extracted instances.

The dataset is archived on ZENODO⁵ and is available at: <https://doi.org/10.5281/zenodo.19256171>.

6. Corpus-based Evaluation

All the experiments are based on the English section of the WWF-LPR corpus as released in TREEN (Pagano et al., 2025).

Each construction is implemented as one or more `Grew` queries, according to the operationalization principles in Section 4. Queries were run independently over the entire corpus. For constructions operationalised using multiple query strategies (such as comparative alternations), matches returned by different queries were combined at the extraction stage under a single construction label. All extraction outputs were exported as structured tables including: (i) construction label, (ii) sentence identifier, (iii) matched token span, (iv) head token, (v) relevant slot assignments.

6.1. Quantitative Measures

To describe how constructional patterns are distributed across the corpus, we combine absolute and normalized frequency measures.

⁵<https://zenodo.org/>

Frequency measures. For each construction, we calculated its raw frequency (total number of hits returned by the query in the corpus) and its normalized frequency, expressed as occurrences per 10,000 tokens, to enable comparison across subcorpora of different sizes.

$$\text{NormFreq}_{c,y} = \frac{N_{c,y}}{T_y} \times 10,000 \quad (1)$$

where $N_{c,y}$ is the number of matches for construction c in year y , and T_y is the total number of tokens in the corresponding subcorpus.

Sentence coverage. We measured the number of unique sentences containing at least one occurrence of a given construction. This indicates how pervasive the construction is in discourse, regardless of how many times it appears in a sentence.

$$\text{Coverage}_{c,y} = \frac{S_{c,y}}{S_y} \quad (2)$$

where $S_{c,y}$ is the number of sentences in year y containing at least one instance of construction c , and S_y is the total number of sentences in the same subcorpus.

Construction density. We also calculated the average number of constructions per sentence to estimate the structural complexity of environmental narrative passages.

$$\text{Density}_y = \frac{M_y}{S_y} \quad (3)$$

where M_y is the total number of construction matches in year y , and S_y is the total number of sentences in the same subcorpus.

6.2. Manual Validation Protocol

To assess the reliability of the operationalization process, we conducted a manual evaluation of the constructions extracted by the automated procedure. All constructions included in the inventory were subject to validation. The aim of this evaluation is not to provide exhaustive annotation of all extracted instances, but rather to estimate extraction precision while capturing variability across construction types.

For each construction, up to five instances per year were randomly sampled from the `Grew` output, resulting in a maximum of 30 instances per construction across the six reports (2014–2024). For constructions with fewer than five instances available in a given year, all available instances were included, leading to a smaller overall sample size.

Sampling was stratified by year and performed over the full set of extracted instances for each construction, ensuring coverage across all subcorpora. Instances were randomly selected from the pool of matches returned by the queries without additional filtering.

Each sampled instance was independently annotated by two annotators, who judged whether the extracted instance corresponded to a valid realization of the target construction. In cases of disagreement, a third annotator adjudicated the final decision.

Full validation results for all evaluated constructions are reported in Appendix A.

Precision estimation. Extraction precision was calculated as the proportion of correctly validated instances over the total number of annotated instances for each construction:

$$\text{Precision} = \frac{C}{N} \quad (4)$$

where C is the number of validated instances and N is the total number of annotated instances for that construction.

Uncertainty estimation. Because the evaluation is based on a limited sample (up to 30 instances per construction), the reported precision values represent estimates rather than full-corpus performance measures. To quantify sampling variability, 95% confidence intervals were computed under a binomial approximation.

The standard error (SE) of the precision estimate is defined as:

$$SE = \sqrt{\frac{p(1-p)}{N}} \quad (5)$$

where p denotes the observed precision. The 95% confidence interval is then obtained as:

$$CI_{95\%} = p \pm 1.96 \cdot SE \quad (6)$$

We emphasize that the evaluation is intended as an initial validation of the operationalization approach rather than a comprehensive assessment of extraction performance. The results nevertheless provide useful indications of the strengths and limitations of query-based construction extraction.

7. Results

This section presents the quantitative and qualitative evaluation of the results of the UD-based operationalization of CASA constructions applied to the WWF-LPR.

7.1. Extraction Precision

In Table 1, we present the precision values obtained from the manual assessment of the sampled constructions. In general, the extraction process demonstrates high reliability for most constructions. For many patterns, we observe high precision values, often around 1.00, for the sampled instances, such as [69] *definite_plural_np*, [68] *definite_singular*, [136] *progressive*, [92] *genitive_np*. A large number of the grammatical patterns which are more central to the descriptive structure of the WWF-LPR show high precision values, e.g. [120] *declarative_statement*, [151] *adjunct*, and [109] *valency_postmodifier*. The observed precision values vary considerably across constructions, reflecting differences in structural ambiguity and the degree of formal constraint in their operationalization. Constructions anchored in stable syntactic dependencies tend to achieve higher precision, whereas more schematic patterns, which rely on broader matching conditions or lexical variation, are more prone to noise. In particular, lower precision values are observed for constructions involving named entities and surface patterns associated with semantically ambiguous configurations, where the interpretation depends more heavily on lexical choice or discourse context. These results highlight the trade-off between coverage and precision in query-based extraction, and underscore the importance of construction-specific tuning.

7.2. Overall Extraction Statistics

Table 2 presents the sentence-level coverage of the selected constructions across the WWF-LPR. In addition to sentence-level coverage, normalized frequency confirms the same tendency. Across all editions, [106] *premodifier_of_noun* emerges as the most pervasive pattern, with coverage values consistently above 0.85 across all years. This stability highlights the strongly nominal and information-dense character of the genre. The consistently high presence of [109] *valency_postmodifier* and related postmodification patterns [107, 108, 113] suggests that noun phrases are not only frequent, but structurally complex, often combining pre- and post-modification within the same syntactic unit. A gradual increase in the use of [66] *indefinite_plural* and [67] *mass_np* constructions also suggests a tendency towards abstraction and systemicity. These constructions have a tendency to refer to collective entities and unbounded conceptual domains, which again emphasizes the macro-level perspective of the reports. Verbal patterns like [120] *declarative_statement* and [3, 4] *monotransitive* are relatively stable across years, showing that even though the reports are assertive and informationally dense, they are not nar-

Table 1: Extract of precision results for selected constructions (eq. 4). N is the number of annotated instances and C the number of instances judged correct. Confidence intervals are computed under a binomial approximation (95%).

Group	Construction	C	N	Precision	95% CI
Best	definite_plural_np	30	30	1.000	[0.884–1.000]
	declarative_statement	29	30	0.967	[0.903–1.000]
	adjunct_base	28	30	0.933	[0.844–1.000]
Intermediate	additive_coordination_base	21	30	0.700	[0.536–0.864]
	monotransitive_clausal_objects	20	30	0.667	[0.498–0.836]
	to_recipient_goal	14	26	0.538	[0.346–0.730]
Lowest	full_name	2	18	0.111	[0.000–0.256]
	on_focus_area	11	30	0.367	[0.194–0.540]
	first_name_np	0	30	0.000	[0.000–0.116]

Table 2: Sentence coverage (eq. 2) of selected constructions in the WWF-LPR (2014–2024).

Construction	2014	2016	2018	2020	2022	2024	avg
premodifier_of_noun	0.848	0.937	0.865	0.885	0.903	0.917	0.893
indefinite_plural	0.673	0.786	0.741	0.716	0.794	0.831	0.757
mass_np	0.502	0.615	0.490	0.657	0.597	0.693	0.592
valency_postmodifier	0.591	0.651	0.721	0.662	0.723	0.661	0.668
declarative_statement	0.494	0.442	0.410	0.397	0.511	0.468	0.454
monotransitive	0.401	0.341	0.347	0.319	0.409	0.421	0.373

rative in nature. Rather, the discourse is organized around general statements, systemic descriptions, and evidence-based claims.

Apart from the high-frequency constructions, a significant part of the data set has zeros or near-zeros coverage across the board for all years. This is not necessarily meaningless data; it can represent genre constraints as well as methodological challenges in operationalizing CASA constructions in UD-based queries. Furthermore, the limited size of the corpus under examination may have influenced the results of the extraction.

First, interactional constructions such as [122] *yes/no questions*, [123, 124] *wh-questions*, and so on are absent from most reports. This is to be expected, given the institutional and expository nature of the WWF-LPR, which prefers informational statements to interaction.

Second, stance markers (e.g., hypothetical or predictive forms) occur sporadically in the corpus (one or two occurrences in particular years), which suggests that epistemic hedging and conditional framing are not important in this genre and are probably restricted to local argumentation. Thirdly, first-person constructions are used sparingly overall, especially first-person singular constructions. Generally, the reports refrain from using first-person constructions in favor of impersonal constructions, nominalizations, and passive constructions. When first-person plural constructions are used, they of-

Table 3: Construction density (eq. 3, matches per sentence) in the WWF-LPR.

Year	M_y	Tokens	Sentences	Density
2014	4156	5373	257	16.17
2016	6253	7655	364	17.18
2018	4529	5870	251	18.04
2020	6251	7777	373	16.76
2022	9449	11864	462	20.45
2024	20114	25806	1006	19.99

ten have institutional rather than personal connotations.

7.3. Construction Density

We also computed the average number of construction matches per sentence ($Density_y$; Eq. 3) to estimate the “density” of constructional information encoded in each sentence in the WWF-LPR subcorpora. In Table 3, we can see that density remains relatively stable across the earlier reports (16.17-18.04 matches per sentence for 2014-2020), and then surges in 2022 (20.45 matches), remaining high in 2024 (19.99 matches).

Nominal density and descriptive packaging. In all years, the preference is strong for dense noun phrases, packaging premodification (for example, *premodifier_of_noun*) with postmodifying complements (for example, *valency_postmodifier*).

(premodification): “environmentally friendly food production”

(postmodification): “the underlying causes of unsustainable development”

Assertion-driven discourse and agency management. The prevalence of *declarative_statement* is coupled with a high incidence of elementary argument-structure patterns (*monotransitive*, [2] *intransitive*) that correspond to the structure of an exposition centered on declarative statements. Passive clauses ([138], *passive_construction*) are less frequent but of great importance, as they tend to focus on the result and the process rather than the actor, especially in cases where they are generic or institutionally diffuse.

(declarative statement): “Human wealth depends on nature’s health.”

(passive): “In this region, climate change is more frequently reported as a threat to populations in the LPI.”

Projection and conditional framing. Modal and hypothetical sentences, although less common, mark instances where the discourse shifts from description to projection. These sentences introduce a reasoning process, which is often conditional, projecting the environmental changes in terms of consequences to be taken into account. The relative infrequency of these sentences indicates that the reports emphasize assertive description over projection.

(hypothetical): “If we lived the lifestyle of a typical resident of the US, we would need 3.9 planets.”

7.4. Error Analysis and Discussion

Although overall extraction performance is consistent, the analysis of mismatches reveals some main sources of errors and limitations, offering useful directions for future work.

UD annotation inconsistencies. A first possible cause of discrepancy relates to the interaction of clause embedding and root assignment in UD. The extraction of *declarative_statement* directly targets finite verbs that are marked as syntactic roots in the sentence. The applicable Grew pattern is:

```
pattern {
  ROOT [];
  V [upos=/^(VERB|AUX)$/, VerbForm="
  Fin"];
  ROOT -[root]-> V;
  S [];
  V -[nsubj|nsubj:pass]-> S;
}
without { V [Mood="Imp"]; }
```

This way, only finite verbs functioning as the matrix root are extracted, excluding subordinate clauses by definition. However, take the following example:

“The result is that we miss the message.”

In the UD representation of this sentence, the verb *miss* is marked as the syntactic *root*, whereas in a canonical UD analysis the clause *that we miss the message* would normally be attached to the copular predicate *is* as a *ccomp*. Consequently, the Grew query correctly extracts this occurrence, since the structural requirement *ROOT-root-V* is met. The problem, therefore, is not in the definition or the logic of the query, but in the representation of the clausal hierarchy as a dependency structure.

Proper names and nominal classification. Another issue is related to proper names and how they combine with constructions like [101] *first_name_np*. The Grew query uses morphosyntactic features like *upos=PROPN* to constrain the search space. However, this feature does not make a distinction between personal names and other named entities. For example:

“The Living Planet Index (LPI) has declined by 52 per cent since 1970.”

In this case, the term *LPI* has the part-of-speech tag *PROPN*, making it relevant to the query, but it does not refer to a personal first name but to an institutional index. Although the extraction is formally correct at the morphosyntactic level, it is semantically not correct for the constructional definition. This could be ideally tackled by enriching queries with external knowledge bases such as WikiData⁶.

Semantically constrained constructions. Some patterns cannot be operationalized purely on the basis of structural constraints (e.g., preposition + noun), but also need some semantic or lexical information. For example, the extraction of the pattern [39, 111] *with_partner* should not be

⁶<https://www.wikidata.org>

formalized solely based on the presence of a prepositional phrase and *with* as the head word, as many other unrelated patterns (instrumental, manner, risk, etc.) would also match. Instead, the pattern should be filtered to include words related to collaboration and partnership (e.g., *partner*, *stakeholder*, *organization*).

(wrong with partner extraction):

“Species are threatened with the risk of extinction.”

Also expressions like [37] *on_focus_area* cannot be analyzed in terms of the structural template *on* + *NP*, but rather require semantic filtering to select cases where *on* expresses topical focus rather than spatial/temporal relations. As already proposed in [Pisciotta et al. 2025](#) in a pilot experiment on Italian, filtering results through WordNet synsets could represent an effective solution.

Structural limits of pattern matching (coordination). Lastly, among the patterns we examined there are constructions sensitive to structural complexity, e.g. additive coordination. The *Grew* patterns usually capture binary coordination configurations; that is, they involve two conjuncts in the coordinated structure. However, there is limited ability to capture more complex coordination chains, with three or more conjuncts. As a result, there is limited ability to capture multi-member coordination.

Lessons learned Constructions based on solid and well-defined dependency relations can be reliably identified in the UD-based operationalization, such as basic noun phrase modification, argument structure constructions like monotransitive and intransitive constructions, and declarative statements. All these exhibit high precision levels. More complex patterns, however, are difficult to handle and hard to model exhaustively in a single query, such as coordinations (especially with more than two conjuncts), long-distance dependencies, and clause embedding, which are more sensitive to root assignment and hierarchical representation in UD. Two classes of constructions emerge: those with a transparent structure in UD that easily correspond to a simple graph configuration, and those without a direct correspondence to a graph. While for the former the link between CASA descriptions and dependency relations is robust and replicable, it is more challenging to extract the latter. Furthermore, annotation inconsistencies and parser artifacts are relevant to constructions that rely on fine-grained morpho-syntactic information. Lastly, semantically restricted constructions, such as partnership or topical focus, cannot be fully expressed using only structural information.

8. Conclusion and Future Work

In this paper, we introduced a novel methodology based on the extraction of constructions and evaluate it on texts about environmental issues. Our main goal is to provide preliminary evidences to confirm the hypothesis that constructions can provide valuable insights into how environmental topics are framed. The methodology consists in semi-automatically convert qualitative construction descriptions, commonly found in constructions such as CASA, into formal graph-based queries in *Grew*. The results of our queries appear to faithfully describe the structure of narratives in a selection of data from the *TREEN* corpus, the first and currently only available treebank for environment, indicating that our approach may be useful for the analysis of environmental discourse.

Our annotation procedure only involved precision: namely, how many hits among the ones matched are true cases of the examined construction. In future work we will expand the methodology to recall as well, trying to spot *false negatives*: cases of construction instances that the *Grew* pattern wasn't able to match. We also intend to apply our methodology on additional datasets of environmental and policy discourse to test its robustness on different datasets and genres. Another line of work will be related to extending our approach to other constructions and checking its portability across languages, examining the potential of CASA-based constructions in UD-annotated corpora of other languages.

Limitations

Several limitations need to be pointed out. Firstly, the precision figures are based on a relatively small manually annotated dataset (up to 30 instances per construction), which is more an approximation than a validation on the entire corpus. Secondly, the approach is dependent on the quality of UD annotation. Inconsistencies in the annotation of dependencies or roots will impact the extraction results, especially for those constructions which are sensitive to the hierarchy or structure of the clauses. Thirdly, the study is limited to the constructions included in the CASA Constructicon. CASA has developed a list of constructions in a structured and linguistically well-motivated way. However, the list is not exhaustive and covers only the whole spectrum of constructions which can possibly play a role in environmental discourse. Lastly, the empirical study is based on a single genre of institutional discourse, WWF-LPR, limiting the immediate generalizability of the study to other types of discourse.

Acknowledgements

The work of Eliana Di Palma was funded by the project NEIKEA (CSP TRAPEZIO Call – Line 1: Paving the Way to Research Excellence and Talent Attraction).

We would like to thank Beatrice Bernasconi for her valuable support and guidance on constructional approaches, which was instrumental in the development of this work.

Bibliography

- Eleni Antonopoulou and Kiki Nikiforidou. 2011. [Construction grammar and conventional discourse: A construction-based approach to discursive incongruity](#). *Journal of Pragmatics*, 43(10):2594–2609.
- Sabine Bartsch, Changxu Duan, Sherry Tan, Elena Volkanovska, and Wolfgang Stille. 2023. [The insightsnet climate change corpus \(iccc\)](#). In *BTW 2023*, pages 887–900. Gesellschaft für Informatik e.V., Bonn.
- Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the University of Turin. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Venice, Italy. CEUR.
- Cristina Bosco, Adriana Silvina Pagano, and Elisa Chierchiello. 2025. [Communicating urgency to prevent environmental damage: insights from a linguistic analysis of the WWF24 multilingual corpus](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 34–43, Tallinn, Estonia. University of Tartu Library.
- Elisa Chierchiello, Patricia Chiril, and Adriana Pagano. 2025. [Ontology-guided domain entity recognition in environmental texts: Evaluating syntax-driven and LLM approaches using BabelNet and GEMET](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 233–244, Cagliari, Italy. CEUR Workshop Proceedings.
- Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, pages 1–54.
- Xu Du, Matthew Kowalski, Aparna S. Varde, Gerard de Melo, and Robert W. Taylor. 2020. Public opinion matters: mining social media text for environmental management. *SIGWEB Newsl.*, 2019(Autumn).
- Charles J. Fillmore. 2008. Border conflicts: FrameNet meets construction grammar. In *Proceedings of the XIII EURALEX International Congress*, pages 49–68, Barcelona. Universitat Pompeu Fabra.
- Charles J. Fillmore, Paul Kay, and Mary Catherine O’Connor. 1988. [Regularity and idiomaticity in grammatical constructions: The case of let alone](#). *Language*, 64(3):501.
- Charles J. Fillmore, Russell R. Lee-Goldman, and Russell Rhomieux. 2012. The FrameNet construction. In Ivan A. Sag and Hans C. Boas, editors, *Sign-Based Construction Grammar*, pages 283–322. CSLI, Stanford.
- Kerstin Fischer and Morgan Aarestrup. 2021. [Relationships between construction grammar\(s\) and genre: Evidence from an analysis of instagram posts](#). *Journal of Pragmatics*, 183:87–104.
- Adele E. Goldberg. 2013. *Constructionist Approaches*. Oxford University Press.
- Francesca Grasso, Stefano Locci, Giovanni Siragusa, and Luigi Di Caro. 2024a. Ecoverse: An annotated twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024*, pages 5461–5472. ELRA and ICCL.
- Francesca Grasso, Ronny Patz, and Manfred Stede. 2024b. NYTAC-CC: A climate change subcorpus based on new york times articles. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa (Italy). CEUR.
- Normunds Gruzitis, Dana Dannélls, Benjamin Lyngfelt, and Aarne Ranta. 2015. [Formalising the Swedish construction in grammatical framework](#). In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 49–56, Beijing, China. Association for Computational Linguistics.
- Bruno Guillaume. 2019. [Graph Matching for Corpora Exploration](#). In *JLC 2019 - 10èmes Journées Internationales de la Linguistique de corpus*, Grenoble, France.
- Thomas Herbst and Thomas Hoffmann. 2018. [Construction grammar for students: A constructionist approach to syntactic analysis \(casa\)](#). *Yearbook of the German Cognitive Linguistics Association*, 6(1):197–218.

- Muhammad Okky Ibrohim, Cristina Bosco, and Valerio Basile. 2023. Sentiment analysis for the natural environment: A systematic review. *ACM Computing Surveys*, 56(4).
- Laura A. Janda, Olga Lyashevskaya, Tore Nessel, Ekaterina Rakhilina, and Francis M. Tyers. 2018. [Chapter 6. A constructicon for russian: filling in the gaps](#). In *Constructicography*, pages 165–182. John Benjamins.
- Zhuo Jing-Schmidt. 2017. [What are they good for? a constructionist account of counterfactuals in ordinary chinese](#). *Journal of Pragmatics*, 113:30–52.
- Knud Lambrecht. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press.
- Adrieli Laviola, Ludmila Lage, Natália Duarte Marção, Tatiane Tavares, Vânia Almeida, Ely Edison Matos, and Tiago Timponi Torrent. 2017. The Brazilian Portuguese Constructicon: Modeling Constructional Inheritance, Frame Evocation and Constraints in FrameNet Brasil. In *Proceedings of the AAAI Spring Symposia*, number SS-17-02 in AAAI Technical Report. AAAI Press.
- Russell Lee-Goldman and Miriam R.L. Petruck. 2018. [Chapter 160.2. the framenet constructicon in action](#). In *Constructicography*, pages 19–40. John Benjamins.
- Per Linell. 2009. *Grammatical constructions in dialogue*, page 97–110. John Benjamins Publishing Company.
- Arthur Lorenzi, Vânia Gomes de Almeida, Ely Edison Matos, and Tiago Timponi Torrent. 2023. [Modeling construction grammar’s way into NLP: Insights from negative results in automatically identifying schematic clausal constructions in Brazilian Portuguese](#). In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 96–109, Washington, D.C. Association for Computational Linguistics.
- Benjamin Lyngfelt, Lars Borin, Kyoko Ohara, and Tiago Timponi Torrent, editors. 2018. *Constructicography: Constructicon development across languages*, volume 22 of *Constructional Approaches to Language*. John Benjamins Publishing Company, Amsterdam.
- Kiki Nikiforidou and Kerstin Fischer. 2015. [On the interaction of constructions with register and genre](#). *Constructions and Frames*, 7(2):137–147.
- Kyoko Hirose Ohara. 2013. [Toward Constructicon Building for Japanese in Japanese FrameNet](#). *Revista Veredas*, 17(1):11.
- Jan-Ola Östman. 2005. *Persuasion as implicit anchoring: The case of collocations*, page 183–212. John Benjamins Publishing Company.
- Adriana Silvina Pagano, Patricia Chiril, Elisa Chierchello, and Cristina Bosco. 2025. [TreEn: A multilingual treebank project on environmental discourse](#). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 80–96, Ljubljana, Slovenia. Association for Computational Linguistics.
- Ludovica Pannitto, Beatrice Bernasconi, Lucia Busso, Flavio Pesciotta, Giulia Rambelli, and Francesca Masini. 2024. Annotating constructions with ud: the experience of the italian constructicon. *arXiv preprint arXiv:2411.07623*.
- Florent Perek and Amanda L. Patten. 2019. [Towards an english constructicon using patterns and frames](#). *International Journal of Corpus Linguistics*, 24(3):354–384.
- Flavio Pesciotta, Ludovica Pannitto, Lucia Busso, Beatrice Bernasconi, and Francesca Masini. 2025. Constraining constructions with wordnet: pros and cons for the semantic annotation of fillers in the italian constructicon. In *Proceedings of the 13th Global Wordnet Conference*, pages 77–84.
- Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact@ACL-IJCNLP 2021*, pages 8–18, Online. ACL.
- Tiago Timponi Torrent, Ely Edison da Silva Matos, Ludmila Lage, Adrieli Laviola, Tatiane Tavares, Vânia Gomes de Almeida, and Natália Sigiliano. 2018. [Towards continuity between the lexicon and the constructicon in FrameNet Brasil](#), page 107–140. John Benjamins Publishing Company.
- Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.
- Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archana Bhatia, Hinrich Schütze, Lori Levin, et al. 2024. Ucxn: Typologically informed annotation of constructions atop universal dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932.

Alexander Ziem and Hans C Boas. 2017. Towards a constructicon for German. In *AAAI Spring Symposium*.

A. Construction List and Validation Results

Table 4: Top constructions extracted from the 2014 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	666	1239.53	84.82
adjunct_base	529	984.55	77.82
adjective_base	450	837.52	77.82
indefinite_plural	393	731.43	67.32
valency_postmodifier	257	478.32	59.14
postmodifier_pp_base	250	465.29	58.37
mass_np	241	448.54	50.19
definite_singular	182	338.73	50.97
declarative_statement	168	312.67	49.42
monotransitive	124	230.78	40.08
third_person_singular_present	121	225.20	38.52
intransitive	109	202.87	35.80
indefinite_singular	92	171.23	27.24
subject_attribute_pp	90	167.50	25.29
asyndetic_coordination	89	165.64	20.62
general_present_tense	85	158.20	25.29
definite_plural_np	55	102.36	17.51
we_definite_reference	55	102.36	16.34
postmodifier_recl	54	100.50	18.68
relative_postmodifier	54	100.50	18.68

Table 5: Top constructions extracted from the 2016 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	1061	1386.02	93.68
adjunct_base	734	958.85	77.20
adjective_base	668	872.63	82.14
indefinite_plural	644	841.28	78.57
mass_np	453	591.77	61.54
valency_postmodifier	407	531.68	65.11
postmodifier_pp_base	405	529.07	65.11
definite_singular	271	354.02	50.00
declarative_statement	206	269.11	44.23
third_person_singular_present	166	216.85	37.91
intransitive	157	205.09	35.44
monotransitive	151	197.26	34.07
asyndetic_coordination	126	164.60	20.60
subject_attribute_pp	126	164.60	26.10
indefinite_singular	121	158.07	27.47
general_present_tense	99	129.33	21.98
definite_plural_np	93	121.49	20.88
monotransitive_clausal_objects	60	78.38	15.11
postmodifier_recl	59	77.07	13.46
relative_postmodifier	59	77.07	13.46

Table 6: Top constructions extracted from the 2018 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	690	1175.47	86.45
adjunct_base	505	860.31	75.30
indefinite_plural	473	805.79	74.10
adjective_base	459	781.94	77.69
valency_postmodifier	363	618.40	72.11
postmodifier_pp_base	361	614.99	71.71
mass_np	227	386.71	49.00
definite_singular	208	354.34	53.78
third_person_singular_present	161	274.28	51.39
declarative_statement	135	229.98	41.04
indefinite_singular	129	219.76	41.43
intransitive	120	204.43	39.44
monotransitive	101	172.06	34.66
asyndetic_coordination	98	166.95	20.32
subject_attribute_pp	71	120.95	23.90
general_present_tense	70	119.25	23.51
definite_plural_np	54	91.99	19.92
passive_construction	54	91.99	18.73
first_name_np	46	78.36	15.14
first_third_past_tense	46	78.36	15.94

Table 7: Top constructions extracted from the 2020 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	1029	1323.13	88.47
indefinite_plural	705	906.52	71.58
adjunct_base	644	828.08	66.49
adjective_base	609	783.08	76.68
valency_postmodifier	477	613.35	66.22
postmodifier_pp_base	475	610.78	66.22
mass_np	429	551.63	65.68
definite_singular	213	273.88	39.95
third_person_singular_present	198	254.60	38.61
declarative_statement	190	244.31	39.68
asyndetic_coordination	158	203.16	20.91
intransitive	145	186.45	31.37
monotransitive	141	181.30	31.90
indefinite_singular	121	155.59	25.20
subject_attribute_pp	109	140.16	22.25
first_third_past_tense	89	114.44	19.03
general_present_tense	82	105.44	19.30
passive_construction	76	97.72	15.82
focus_np	74	95.15	19.84
first_name_np	70	90.01	13.94

Table 8: Top constructions extracted from the 2022 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	1478	1245.79	90.26
adjunct_base	1100	927.17	81.82
indefinite_plural	1037	874.07	79.44
adjective_base	881	742.58	82.03
valency_postmodifier	716	603.51	72.29
postmodifier_pp_base	714	601.82	71.86
mass_np	547	461.06	59.74
definite_singular	358	301.75	52.81
declarative_statement	300	252.87	51.08
third_person_singular_present	279	235.17	47.62
intransitive	232	195.55	41.56
monotransitive	218	183.75	40.91
indefinite_singular	208	175.32	35.93
asyndetic_coordination	204	171.95	20.56
subject_attribute_pp	165	139.08	27.49
general_present_tense	164	138.23	30.09
first_name_np	132	111.26	19.48
last_name	132	111.26	19.48
definite_plural_np	117	98.62	22.51
passive_construction	109	91.87	20.78

Table 9: Top constructions extracted from the 2024 WWF Living Planet Report. Frequency is normalized per 10,000 tokens.

Construction	Matches	Freq/10k	Sent. Coverage (%)
premodifier_of_noun	3446	1335.35	91.65
indefinite_plural	2405	931.95	83.10
adjunct_base	2348	909.87	81.01
adjective_base	2009	778.50	80.52
mass_np	1633	632.80	69.28
valency_postmodifier	1257	487.10	66.10
postmodifier_pp_base	1250	484.38	65.90
definite_singular	711	275.52	47.02
declarative_statement	617	239.09	46.82
third_person_singular_present	543	210.42	40.26
monotransitive	508	196.85	42.15
asyndetic_coordination	500	193.75	28.33
intransitive	457	177.09	36.68
indefinite_singular	363	140.66	29.82
general_present_tense	337	130.59	27.93
subject_attribute_pp	313	121.29	23.06
definite_plural_np	246	95.33	20.87
monotransitive_clausal_objects	202	78.28	17.79
postmodifier_recl	197	76.34	17.20
relative_postmodifier	197	76.34	17.20

Table 10: Full validation results for all constructions, reported using the same precision metric as in Table 1.

Construction	C	N	Precision
first_name_np	0	30	0.00
last_name	0	30	0.00
on_focus_area	11	30	0.37
for_desired_object	12	30	0.40
subject_attribute_pp	14	30	0.47
monotransitive_clausal_objects	20	30	0.67
additive_coordination_base	21	30	0.70
postmodifier_acl_partger	21	30	0.70
with_partner	21	30	0.70
general_postmodifier_of_adjective	22	30	0.73
indefinite_plural	23	30	0.77
relative_postmodifier	25	30	0.83
intransitive	26	30	0.87
premodifier_of_adjective	26	30	0.87
focus_np	27	30	0.90
postmodifier_pp_base	27	30	0.90
adjunct_base	28	30	0.93
mass_np	28	30	0.93
valency_postmodifier	28	30	0.93
adjective_base	29	30	0.97
alternative_coordination	29	30	0.97
asyndetic_coordination	29	30	0.97
declarative_statement	29	30	0.97
monotransitive	29	30	0.97
numerical_plural	29	30	0.97
passive_construction	29	30	0.97
premodifier_of_noun	29	30	0.97

Construction	C	N	Precision
definite_plural_np	30	30	1.00
definite_singular	30	30	1.00
demonstrative_plural	30	30	1.00
demonstrative_singular	30	30	1.00
general_present_tense	30	30	1.00
general_quantifier_singular	30	30	1.00
genitive_np	30	30	1.00
indefinite_singular	30	30	1.00
it_definite_reference	30	30	1.00
most_least	30	30	1.00
perfective_construction	30	30	1.00
postmodifier_recl	30	30	1.00
progressive	30	30	1.00
superlative	30	30	1.00
third_person_singular_present	30	30	1.00
us_definite_reference	30	30	1.00
we_definite_reference	30	30	1.00
first_third_past_tense	29	29	1.00
they_definite_reference	29	29	1.00
ranking_np	26	28	0.93
general_quantifier_plural	28	28	1.00
to_recipient_goal	14	26	0.54
them_definite_reference	26	26	1.00
self_motion	17	25	0.68
general_past_plurals	24	24	1.00
general_quantifier_mass	24	24	1.00
complex_full_name	20	23	0.87
ditransitive_objtoinf	20	21	0.95
imperative_construction	21	21	1.00
self_change_resultative_base	14	19	0.74
full_name	2	18	0.11
attribute_construction_adj	10	18	0.56
x_is_adj_to	16	17	0.94
subject_attribute_adjp	2	15	0.13
caused_motion	12	14	0.86
adjectival_with_partner_base	4	10	0.40
ditransitive_obj_inf	7	10	0.70
realistic_condition	7	10	0.70
comparative_than	10	10	1.00
ditransitive_objnp	8	8	1.00
we_are_able_to_make	8	8	1.00
x_is_likely_to	8	8	1.00
itself_reflexive	1	7	0.14
change_of_state	0	6	0.00
caused_change_resultative	6	6	1.00
hypothetical	6	6	1.00
to_recipient_for_desired	1	4	0.25
about_reference_area	4	4	1.00
you_definite_reference	4	4	1.00
wh-cleft	0	2	0.00
ourselves_reflexive	2	2	1.00
ditransitive_wh-cl	1	1	1.00
exchange_x_for_y	1	1	1.00
first_person_singular_present	1	1	1.00
i-definite_np	1	1	1.00
it_cleft	1	1	1.00

Construction	C	N	Precision
resultative_attradj	1	1	1.00
wh-question_subject	1	1	1.00
with_emotion	1	1	1.00
yes_no_question	1	1	1.00

Mapping the Historical Ecology of the Cyclades: A Diachronic Natural Language Processing Analysis of Travel Narratives (1700–1920)

Aikaterini Christopoulou^{1,2,3,4}, Vassilis Detsis¹, Basilis Gatos^{2,3,4}

¹Harokopio University of Athens, ²National Centre for Scientific Research “Demokritos”,

³AI4DOC P.C., ⁴Archimedes Unit - Athena Research Centre

¹El. Venizelou Ave. 70 17676 Kallithea Greece, ^{2,3}Patr. Gregoriou E & 27 Neapoleos Str 15341 Agia Paraskevi Greece, ⁴Artemidos Street 1 15125 Marousi Greece
achristopoulou@hua.gr, detsis@hua.gr, bgat@iit.demokritos.gr

Abstract

Historical texts can be valuable for the study of a place's ecological history but reading and extracting information from them can be a tedious and time-consuming task. Natural Language Processing can help in order to extract the most important information of the text in a quick, effective and reproducible way. In this study, travel narratives for the Cyclades Islands from 4 different time periods (1700-1920) have been chosen for analysis. The first step, the quantitative part, includes the semi-automatic detection of geographical entities in the texts and their connection to predefined keywords in order to enable temporal and spatial statistical analysis. The output of this procedure is then inserted in a Retrieval-Augmented Generative Synthesis pipeline in which the text segments with the connected place and keyword are processed by a locally orchestrated Large Language Model. The final output is used for the understanding and interpretation of the original text. Even though the study focuses mainly on the coherence and repeatability of the workflow, an effort is made to interpret and connect the results to the past ecological profile of these islands. The dataset/supplementary material is provided via an open access repository.

Keywords: Name Entity Recognition, Large Language Model, Historic Narratives

1. Introduction

Travelers of the 17th to 20th century left an invaluable legacy of historic texts which they wrote in their voyage journals. Some of these travellers who have travelled in Greece during this time are William Martin Leake, Sir William Gell, François Pouqueville, James Theodore Bent and Joseph Pitton de Tournefort. Either archaeologists, historians, geographers, botanists, scholars or even militaries, all of them had a passion for discovery, using their sharp observation and keen curiosity to understand, interpret and transfer landscape and events in such a way, that even 200-400 years later, their texts can be object of academic research. In most cases their journeys included mainland Greece, e.g., Attika and the Peloponnese, and insular Greece, e.g., the Cyclades and Crete Island. In their texts they included various information about the places they visited, such as the description of landscape features, fauna and flora, as well as, accounts of activities, folklore and the way of life of the inhabitants of these places, attributed with each author's personal style. Thus, these texts are well suited for the historic study of past landscapes, both from an ecological and a linguistic point of view. Their value lies in the fact that they are some of the very few depictions of what the Greek territory was like in this time, especially during the Greek occupation by the Ottomans, as well as the 19th and early 20th century.

Despite of their value and importance, historical texts of travellers have been overlooked especially by geographers and ecologists, while

in fact, they can be used for the evaluation of historic ecosystem services' provision and landscape reconstructions (Tomscha et al., 2016). They might be scarce and very few in number, but they are still lengthy and dense texts, whose analysis can be a time-consuming task if done manually. The use of modern technologies for textual understanding and analysis, also known as Natural Language Processing (henceforth NLP), can be a very useful and efficient solution to this task. NLP can be used for a quick, systematic and structured screening of large volumes of such texts, in order for the researcher to (1) get the general idea of each narrative, (2) sort out or prioritise the most important, rich and promising sections, (3) understand the views and personal style of the writer and, most importantly, (4) achieve all of the above in a repeatable and reproducible way.

Although NLP has been successfully applied in various linguistic and environmental research works, it has never, to our knowledge, been applied in historic texts of travellers, especially, with the focus on ecological research. Thus, this study aims to bridge the gap between these historical narratives, temporal and spatial ecological research and current state-of-the-art text understanding methods. The goal of this work is the quantitative and qualitative analysis of 4 historic travellers' texts, from 4 different time periods (1700, 1830, 1880 and 1920), using NLP tools in R programming environment. The quantitative analysis was implemented using Name Entity Recognition (NER) techniques and the qualitative analysis was carried out by using local Retrieval-Augmented Generation (RAG)

Large Language Models (henceforth LLMs). The 4 texts that have been chosen for the analysis, supplementary material and results (tables and high-resolution figures) are included in an open access dataset/supplementary material provided¹ and refer to Cyclades as a study area. Each Cycladic Island holds its own unique physical, natural and environmental characteristics, which lead to equally unique activities, production and way of life of its inhabitants. Additionally, they have a strategic location as they are part of the maritime route that connects mainland Greece to Asia Minor, Constantinople and the Holy Land. The Cyclades is one of the few examples of Greek areas that remained more or less the same during the turbulent times of the 17th to 19th century, making them a living laboratory, suitable for the study of landscape and ecological change and human-environment interaction impacts. Due to all of the above, the texts that refer to the Cyclades were considered a more suitable case study for this kind of analysis. Even though the study focuses mainly on the coherence and repeatability of the workflow, an effort is made to interpret and connect the results to the past ecological profile of these islands.

2. Related Work

The interest in the use of Natural Language Processing (NLP) among several disciplines of environmental studies is gradually increasing. The subjects that have been assessed using NLP methods include landscape or biodiversity (Schimanski et al., 2023; Abdelmageed et al., 2022; D'Souza et al., 2025), climate change (Grasso and Locci, 2024), ecosystem services provision (Havinga et al., 2024; Kong et al., 2023; Luo et al., 2025; Zhang et al., 2026), natural hazards (Avcioglu et al., 2025; Lai et al., 2022; Sodge et al., 2023) and sustainability-related regulations (Villacampa-Porta et al., 2025). The majority of these studies use as data sources either social media content (Havinga et al., 2024) or modern newspaper and journal articles (Avcioglu et al., 2025; Sodge et al., 2023). Current works can also be distinguished in three distinct categories, those that use mainly statistical methods, such as Name Entity Recognition (Abdelmageed et al., 2022), those that integrate Large Language Models in their workflow (Zhu et al., 2023; Grasso and Locci, 2024; Luo et al., 2025; Zhang et al., 2026) and those that combine the aforementioned methods (Nundloll et al., 2022). Noteworthy, there are numerous studies that incorporate geographic or spatial features in their pipeline, showcasing the opportunities for connection and interdisciplinarity between geographic, linguistic and environmental subjects (e.g., Avcioglu et al., 2025; Gregory et

al., 2015; Grossner et al., 2026). What is also worth mentioning is a strong connection between sentiment analysis and environmental studies (Havinga et al., 2024; Barz et al., 2025; Huai and Van De Voorde, 2022). However, the studies that involve the analysis of historical data, such as censuses (e.g., Haider et al., 2025), records (e.g., Li and Shi, 2025) or texts (e.g., Nundloll et al., 2022; Brando and Frontini, 2017) are scarce. Thus, our work focuses on historical travel accounts, as they combine temporal and spatial scale, needed in ecological studies, accompanied by the personal writing and narrative style, which can influence the text understanding and interpretation.

3. Dataset

The corpus is divided into 4 parts as it consists of 4 seminal travelogues covering most of the Cyclades (e.g. Amorgos, Andros, Melos, Naxos, Santorini and Syros) and dated at 1700, 1830, 1880, and 1920.

The first and oldest part, includes text transcriptions from 233 pages of the book «A voyage into the Levant», translated from the original French book «Relation d'un voyage du Levant»², by Joseph Pitton de Tournefort, a French doctor, botanist and traveler. The book consists of 3 volumes and it was published in 1741, in London. These texts have some peculiarities that need to be taken into consideration. The first involves the use of older place names some of which have no nomenclatural connection to the modern name of the island. The second is the use of the archaic long s (shown as «f» or «ff» in the text). The third refers to the language itself where in most cases words are in the form of «parch'd». All the above make the texts difficult for an LLM to read, understand and analyse.

The second part includes text transcriptions from 21 pages of the book «Observations Upon the Peloponnesus and Greek Islands»³, by Rufus Anderson, an American minister. The book was written in 1829 and it was published in Boston in 1830. Anderson travelled in the Peloponnesus, Aegean and Ionian islands, gathering information concerning the geography, history and culture of the places he visited, reflecting in social and political issues of the time. As his main mission during his overseas travel was the spread of Christianity, the main theme of this book is occupied by matters of faith and the Greek church.

The third part includes text transcriptions from 321 pages of the book «The Cyclades, or; Life among the insular Greeks»⁴, by James Theodore Bent an

¹ Dataset/Supplementary material accessed from the link: <https://zenodo.org/records/19226329>.

² <https://catalog.hathitrust.org/Record/001240678>

³ <https://catalog.hathitrust.org/Record/000649704>

⁴ <https://hdl.handle.net/2027/mdp.39015028327800>

English explorer, archaeologist, and author. The book was published in London in 1885 and it is a personal account of two tours made by Bent and his wife in the Greek islands between 1882 and 1884. It has been characterised as a classic guide to the Cyclades. His narrative includes a very personal, quirky tone, as he describes the various pleasant and unpleasant events while visiting the Cyclades, describing the landscape, social life, production and activities in a detailed manner.

The fourth part includes text transcriptions from 80 pages of the book «A Handbook of Greece»⁵, which was published in 2 volumes in London in 1919 and it was compiled by the Geographical Section of the Naval Intelligence Division, Naval Staff, Admiralty. The selected pages include text that refer to physical, natural and environmental characteristics of the Cyclades in general, as well as social and economic aspects of the Islands. This book was added to the corpus as it is more scientific and data-driven, in comparison, especially, with the journals of Tournefort and Bent which are more narrative, descriptive and even poetic.

<p>(a) threatens. We therefore resolv'd to wait for a French Bark: by good luck there was at Canea one of those which your Lordship has forbid pickeering from Island to Island for Plunder. I promis'd the Master not to inform againft him, and fo he convey'd us to Argentiere, the first of August. ΚΙΜΩΛΟΣ. This Island, by the Greeks call'd Chimoli [b], took the name of Argentiere at the time when the Silver Mines were first discover'd there: there are still to be seen the Work-houses</p>
<p>(b) We therefore resolv'd to wait for a French Bark: by good luck there was at Canea one of those which your Lordship has forbid pickeering from Island to Island for Plunder. I promis'd the Master not to inform againft him, and fo he convey'd us to Argentiere, the first of August. This Island, by the Greeks call'd Chimoli [b], took the name of Argentiere at the time when the Silver Mines were first discover'd there</p>
<p>(c) We therefore resolv'd to wait for a French Bark: by good luck there was at Canea one of those which your Lordship has forbid pickeering from Island to Island for Plunder. I promis'd the Master not to inform against him, and so he convey'd us to Argentiere, the first of August. This Island, by the Greeks call'd Chimoli [b], took the name of Argentiere at the time when the Silver Mines were first discover'd there</p>

Table 1: A sample of the oldest book of the corpus «A voyage into the Levant», (a) original text, (b) the result of the automatic OCR and (c) OCR corrections marked in red.

⁵ <https://catalog.hathitrust.org/Record/008881832>

4. Methodology

This study introduces a multi-phase Natural Language Processing (NLP) framework which was designed (a) to transform unstructured historical journals of travellers of the 17th to 20th century into structured data relevant for ecological studies and (b) to analyse and semantically interpret the ecology of the landscape. The framework integrates statistical text mining with retrieval-augmented neural generation, enabling both quantitative pattern detection and qualitative interpretation. The proposed architecture integrates rule-guided linguistic annotation, lexicon-based semantic extraction, and neural generative modelling (Figure 1). The pipeline consists of two sequential components: (a) A statistical information extraction module which performs linguistic preprocessing name entity identification, lexicon-guided semantic matching and context-based relationship extraction. (b) An LLM based retrieval-augmented neural synthesis module which generates semantically coherent ecological descriptions grounded in retrieved textual evidence. The framework follows a corpus-driven paradigm, combining deterministic extraction with neural interpretation while preserving traceability between generated outputs and source text.

4.1 Data Preparation and Lexicon Construction

The corpus consists of digitized historical travel journals containing descriptions of landscapes, ecological conditions, activities, products and the way of life. Given the linguistic variability and orthographic inconsistencies typical of historical documents, the oldest text (1700) was corrected concerning the automatic Optical Character Recognition (OCR) which was originally provided (Table 1) and all texts were normalized through a preprocessing stage including tokenization, sentence segmentation, and lemmatization. To guide the extraction process, a domain-specific ecological lexicon was constructed comprising 130 keywords organized into thematic categories such as activities (e.g., mining, pottery), harvest (e.g., wheat, olive), products (e.g., wine, jam), infrastructure (e.g., harbour, bridge), natural resource (e.g., marble, emery), water (e.g., stream, river) and landscape (e.g., grove, cave). The lexicon was developed through expert-driven selection based on ecological relevance and expected occurrence in historical descriptive narratives. This lexicon serves as the primary semantic anchor for subsequent statistical and generative analysis.

4.2 Statistical Keyword Pipeline

The first phase of the framework performs statistical extraction of ecological information through entity identification and proximity-based

relationship analysis. This phase combines automated linguistic annotation with expert-guided refinement to ensure accuracy in historical contexts characterized by spelling variation, archaic syntax, and toponymic evolution.

4.2.1 Entity Refinement

Geographical entities were initially detected using part-of-speech tagging and dependency parsing (Wijffels et al., 2018). However, historical corpora present unique challenges, including obsolete place names, spelling variations, and ambiguous entity references. To address these challenges, a semi-automated Human-in-the-Loop refinement stage was introduced. The automatically extracted location candidates were reviewed and curated by domain experts to resolve toponymic ambiguities, normalize historical place names to standardized equivalents (e.g., mapping “Argentiere” to its modern equivalent “Kimolos”) and remove false positives. This process resulted in a validated gazetteer of geographical entities, which was subsequently used as a reliable reference for downstream relationship extraction and retrieval operations. This hybrid human-machine approach significantly improves entity precision while preserving scalability.

4.2.2 Proximity-Based Relationship Extraction

Following entity validation, the framework identifies semantic relationships between geographical locations and ecological concepts using a sliding window co-occurrence algorithm (Silge and Robinson, 2016). This approach captures both explicit and implicit associations within narrative contexts. Two complementary extraction strategies were employed:

4.2.2.1 State-Persistent Contextual Tagging

This strategy models the narrative flow of the text by propagating the most recently mentioned geographical entity across subsequent tokens until a new location is encountered. Ecological keywords appearing within this propagated context are associated with the active spatial anchor. This method captures extended descriptive passages in which ecological features are discussed without repeated explicit mention of the location.

Formally, given a sequence of tokens t_1, t_2, \dots, t_n , and detected location tokens $L_i \subset T$, each ecological keyword occurrence k_j is assigned to the nearest preceding location entity L_i , provided no intervening location entity exists. This approach enables reconstruction of narrative-level geographic-ecological associations.

4.2.2.2 Keyword-in-Context (KWIC) Extraction

To ensure high-precision extraction suitable for downstream neural processing, a localized context window approach was also applied. For each detected geographical entity (e.g., Andros), a symmetric window of ± 20 tokens, was

extracted. These contextual segments were scanned for ecological keyword occurrences (e.g., olive, silk) using boundary-aware lexical matching. This method produces high-confidence, localized evidence of ecological-geographical relationships (e.g., consistent presence of silk in 1700 and 1880 in Andros Island) and serves as the primary evidence source for retrieval-augmented generation.

4.3 Retrieval-Augmented Generative Synthesis Pipeline

While statistical keyword extraction effectively identifies candidate relationships, it cannot fully capture the semantic richness and implicit ecological descriptions present in historical narratives. To address this limitation, the second stage employs a Retrieval-Augmented Generation (RAG) framework that combines evidence retrieval with neural language modeling. The output of the statistical pipeline is transformed into a structured knowledge index containing temporal metadata, ecological keyword and contextual text segment. Rather than processing the entire corpus, the system selectively retrieves high-density textual segments where ecological and geographical entities co-occur. This targeted retrieval improves computational efficiency while maximizing semantic relevance. Each retrieved segment serves as grounded input for generative interpretation.

The retrieved evidence segments are processed by a locally orchestrated Large Language Model (Llama 3) to generate structured ecological descriptions. The model was prompted to synthesize information across multiple textual fragments, producing coherent interpretations of historical landscapes focusing on human-environment interactions. Unlike purely generative approaches, the model operates within a constrained retrieval-augmented setting, ensuring that all generated content is grounded in verifiable textual evidence rather than unconstrained inference. An example of a prompt provided is this: “Analyze this historical snippet mentioning Andros. Extract four ecological data points in exactly this format, separated by pipes: LANDSCAPE | EXPLOITATION | WATER | PRODUCTION. Consider as mentioned if 1-4 words are included for each category. If not mentioned, write 'NA'”.

4.4 Implementation Details

The proposed pipeline was implemented using a modular architecture designed to support corpus preprocessing, linguistic annotation, lexicon-based information extraction, and retrieval-augmented neural generation. The implementation was developed in the R programming environment, enabling reproducible corpus processing and seamless integration of statistical and neural components.

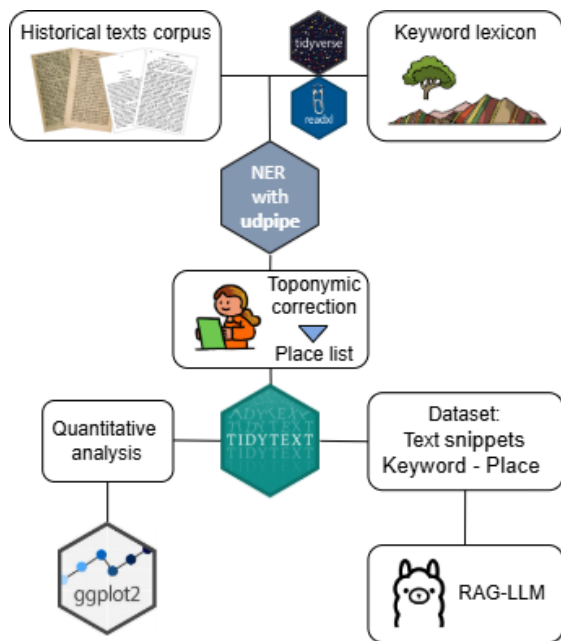


Figure 1: Hybrid NLP Framework for Historical Ecology Analysis.

4.4.1 Linguistic Annotation and Corpus Processing

Linguistic preprocessing, including tokenization, sentence segmentation, lemmatization, and morphosyntactic annotation, was performed using the **udpipe** package (Wijffels et al., 2018). This framework implements models based on the Universal Dependencies formalism, providing standardized annotations including lemma normalization, part-of-speech tagging, morphological feature annotation and dependency parsing. These annotations enabled robust lexical matching and improved resilience to orthographic and morphological variation present in historical texts. The annotated output was stored in structured tabular format, allowing efficient downstream querying and contextual extraction.

4.4.2 Lexicon-Based Extraction and Contextual Analysis

Lexicon-driven ecological concept detection and contextual co-occurrence analysis were implemented using the **tidytext** package (Silge and Robinson, 2016) which provides corpus-oriented text mining functionality based on token-level data representations. The tidytext framework enabled token-level corpus representation using tidy data principles, efficient keyword matching at the lemma level, implementation of sliding window context extraction, Keyword-in-Context (KWIC) analysis and context propagation for entity-keyword association. Regular expression matching with boundary constraints was used to

ensure precise detection of ecological terms while avoiding spurious partial matches.

Corpus transformation, filtering, and aggregation operations were implemented using the **tidyverse** package (Wickham et al., 2019b), which provides efficient and reproducible data manipulation capabilities. These operations included corpus normalization, token filtering and aggregation, context window construction, knowledge base assembly.

4.4.3 Gazetteer Curation and External Resource Integration

Supplementary structured resources, including curated gazetteers and lexicon files, were managed using the **readxl** package (Wickham et al., 2019a). This enabled integration of manually curated entity normalization mappings into the automated pipeline. The curated gazetteer was used to normalize historical toponyms, resolve spelling variation and support consistent entity linking. This ensured high precision in geographic entity resolution.

4.4.4 Retrieval-Augmented Neural Generation

The Retrieval-Augmented Generation (RAG) component was implemented using the **ollamar** framework (Lin and Safi, 2024), which provides an interface for local deployment and orchestration of transformer-based large language models. The system performs the following operations: retrieval of contextually relevant evidence segments from the knowledge base, construction of structured prompts combining geographic and ecological information, neural generation of evidence-grounded ecological interpretations. This local deployment approach ensures reproducibility, data privacy, reduced computational latency and full control over model inference.

4.4.5 Visualization and Statistical Analysis

Statistical summaries and visualizations of extracted ecological patterns were generated using the **ggplot2** package (Wickham, 2016), which implements a grammar-of-graphics approach. Visualization outputs included temporal frequency distributions of ecological keywords, geographic distributions of ecological references and co-occurrence frequency plots. These visualizations support exploratory corpus analysis and facilitate interpretation of extracted ecological patterns.

5. Results and Discussions

5.1 Quantitative Analysis: The Keyword Pipeline

The comparison of keyword occurrences belonging in the categories “Landscape”, “Activities”, “Inland Water” and “Infrastructure”, between the 4 time periods, is presented in Figure

2. The description of “Landscape” features is dominant in every text included in the analysis, while keywords describing “Activities” are the least mentioned, presenting the highest frequency

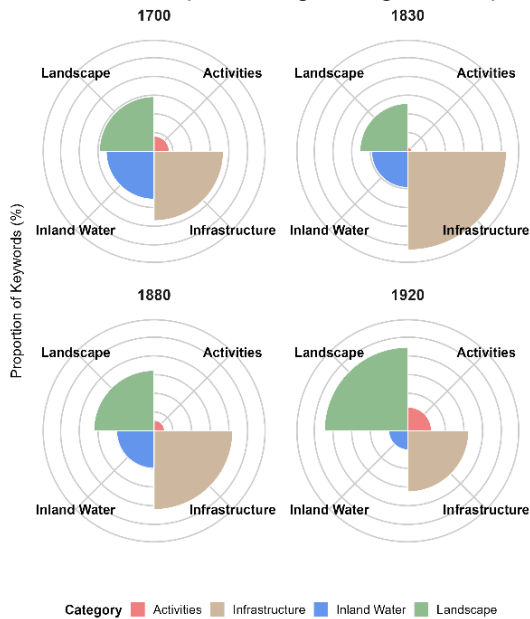


Figure 2: A macro-level thematic context of the keyword-island connection analysis.

in the text of 1920. “Inland Water” shows a gradual decrease, while “Infrastructure” peaks in the 1830 text. The decrease in “Inland Water” occurrences could be probably attributed to the absence of the specific keywords included in the lexicon rather than an actual unavailability of water. These could

be counted as false negatives, as the absence of keywords belonging to a category e.g., “Inland Water”, does not necessarily mean the actual lack of availability of water. This could be validated by the results of the LLM analysis (see Section 5.2 and Figure 5).

In Figure 3, a comparison between the density of specific keywords, for 1700 and 1880, is presented. The narrative of 1700 focuses heavily on the presence of “Product” or “Harvest” (high frequency for wine, oil, cotton, etc.), while the narrative of 1880 shows a surge in “Activities” (pottery or mines). This suggests a transition from a consumer/observer perspective, to a more industrial or ethnographic perspective of documenting people’s way of life. The “Product” category undergoes a clear evolution in variety. While in 1700 the narrative was dominated by “wine”, “oil” and “silk”, in 1880 “wine” maintains a high frequency, but the emergence of processed goods like “honey”, “cheese” and “jam” is observed. This could indicate a more sophisticated economic reporting style or an actual diversification of the Aegean export economy during the 19th century.

In Figure 3, the keyword “mines” remains consistent, but its location shifts or intensifies in specific areas like Melos and Naxos across the two periods. A high frequency of occurrences of specific words, e.g., “rock” and “flower” for the text of 1700 or “mountain”, “hill” and “ground” for the text of 1880 can be observed. In 1700, keywords are clustered around specific islands like Delos,



Figure 3: Keyword frequency comparison between 1700 and 1880 for 6 different categories in all the islands included in the texts.

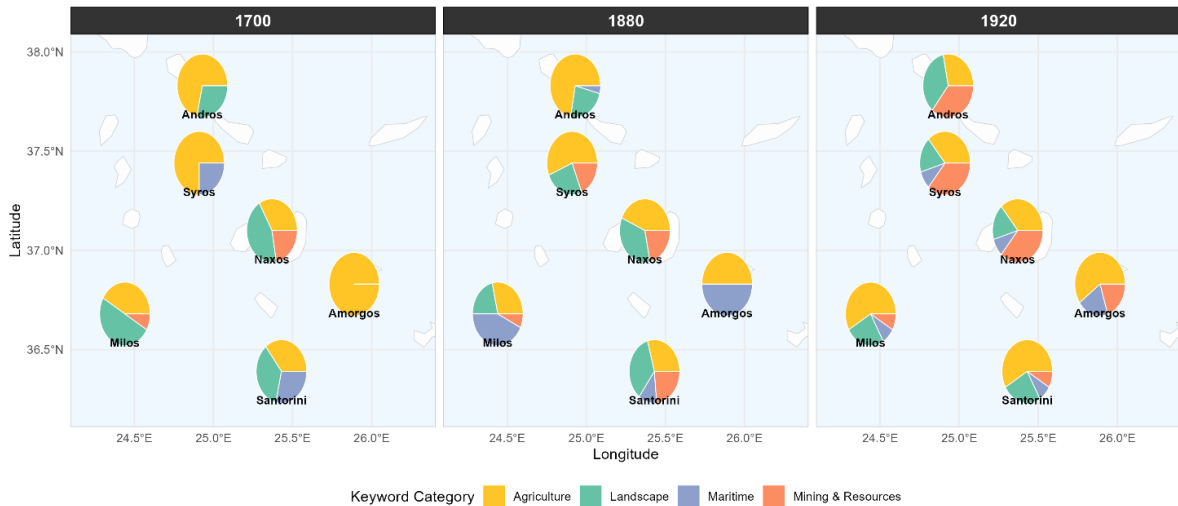


Figure 4: Spatial distribution of 4 ecological themes. Pie charts represent the relative frequency (mentions per 1,000 words) of keywords across 3 historical journals.

Melos, and Naxos. In 1880, the distribution is much more spread across smaller islands like Amorgos, Anaphi, and Keos. Also, there is a noticeable emergence of “cave” mentions in 1880 (specifically in Andros and Antiparos) that barely existed in 1700.

The difference in the text contents can be mainly attributed to the author’s interest and point of view and not at the presence or absence of specific features in the landscape, e.g., it is impossible that hills or ground to be absent from the landscape in 1700. Another reason for the absence of specific features, e.g., “cave”, could be attributed to the difficulty of access that would characterize the journey of a traveller in comparison to a more recent one.

There is a persistent occurrence of specific words across all places and time periods, e.g. the word “wine” or “rock” in Figure 3. These are attributed in most cases to the objective truth, as the Cyclades are, indeed, areas with rocky landscapes with scarce vegetation, especially during the summer. In cases like these, one can export safe conclusions, e.g., for the continuous production or trade of wine across the selected time periods.

There is high density of words for a specific time period or place, e.g. the word “plant” for the 1700’s text or “mountain” for the 1880’s text (Figure 3). These can be attributed to the personal interests of the author of the specific text, as more specifically, Tournefort was a botanist, so his interest in plants justifies the high density of the word in his texts, more than drawing the assumptions that plants had a higher occurrence in the Cyclades in 1700 than in 1880. This can also be observed in Figure 2 where the highest occurrences of keywords belonging in the category “Infrastructure” was attributed in the text of 1830, presumably because of the profession of

the author as he was a minister with an interest in religion and education. Thus, occurrences like these cannot be used as evidence for changes, in the absence of further indications.

In Figure 3, there are scarce, low in frequency word occurrences, such as the “corn”, “barley” and “mulberry”. These are the cases that need to be cross-validated with additional historical sources. For example, according to Kolovos (2007), Andros was one of the highest silk producers in Greece the 17th and 18th century, thus the connection of the word with this island is justified. The presence of other agricultural products in the Cycladic Islands can, additionally, be validated with agricultural censuses, the oldest of which was in 1911. Another interesting finding is the striking absence of “figs” from the 1880’s texts which could, also, be validated from Kolovos (2017), who quoted Horden and Purcell (2000) who suggested that «there is no other edible fruit in the Mediterranean that was so unfairly abandoned like figs».

Figure 4 which includes the granular island profiles (Amorgos, Andros, Melos, Naxos, Santorini and Syros) reveals distinct ecological and economic trajectories. A major “mining shift” is observed in all 6 islands. While mining was present only in Naxos and Melos Islands during the 17th century, the occurrences of words related to mining and natural resources increase drastically during the 20th century texts. A more stable presence of “Agriculture”-related keywords, during every period reflects the agrarian focus especially of early travellers. More specifically, Amorgos Island shows a clear transition from 1700’s agricultural focus to an interest in mining and maritime activity in 1880. For Andros Island, the data captures the consistent presence of silk in 1700 and 1880 (Figure 3) and a continuation of agricultural activities and production even in 1920. Notably, it

also records a sudden shift toward industrial mineral extraction in 1920. All these changes occurring in Andros Island justify the focus on this island for the LLM analysis pipeline. As for Naxos Island, it demonstrates long-term agricultural stability and it is the only island with a steady presence of mining activity throughout the studied years.

5.2 Qualitative Synthesis: The LLM-RAG Pipeline

The use of the Retrieval-Augmented Generation (RAG) component in the ollama framework, assisted in the understanding and interpretation especially of those aspects that were either overlooked or misinterpreted by the keyword-spotting pipeline. A total of 130 snippets, that were produced from the previous phase and refer to Andros Island, were used as an input to the

the LLM, seen in Figure 5, does not imply such a decrease.

6. Conclusion

The combination of the keyword pipeline and the LLM-RAG pipeline is a holistic approach towards understanding and analysis of old text contents, whose study can be very useful in the field of historical ecology but very challenging and time-consuming when done manually. The keyword pipeline provides the statistical validity and longitudinal scope, while the RAG pipeline provides the qualitative depth necessary to understand human-environment interactions. This dual approach mitigates the risk of “keyword spotting” without context, while also preventing the “hallucination” risks associated with running LLMs on unfiltered, massive text volumes.

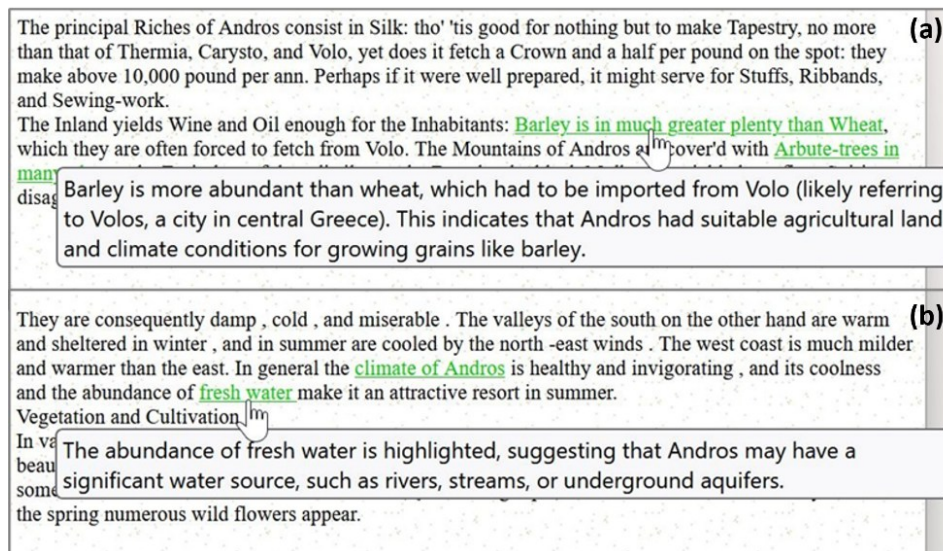


Figure 5: The LLM output incorporated in the original text of (a) 1700 and (b) 1920.

LLM framework. Figure 5 shows an example from the enriched text output containing the LLM result.

The first important issue that the LLM handles well is the competing entities in the same sentence like in the case of the occurrences of the words “barley” and “wheat” in the sentence «Barley is in much greater plenty than Wheat», which is included in the input dataset to the RAG model. The occurrence of these words is counted as equal when it comes to keyword frequencies but according to the aforementioned sentence, the context of comparison is missing. A second issue that was handled with the LLM, refers to false positives of the first phase, e.g. “silk” in Andros Island, that was detected in the sentence «the silk industry, once flourishing, has now almost disappeared from Andros» of the 1920 text. A third example of the disagreement between the two phases has to do with “water”. According to Figure 2, the category “Inland Water” shows a gradual decrease in team, however, the output of

6.1 Limitations

Despite of the usefulness and repeatability of the indicated methodology there are some limitations that should be addressed. The first refers to the construction of the keyword lexicon, which in the present study was compiled independently of the corpus, leading to possible omissions of keywords, especially when it comes to older texts. A wider, more appropriate and more focused keyword lexicon can lead to a deeper insight and understanding of the texts. A second limitation refers to the content of the text, as in many cases it is the result of personal perspective or “agenda” of the author. This is addressed by the evaluation with additional historical texts and published works. A third limitation involves the limited size of the dataset used for this analysis, with the results being influenced by the authors’ personal views. Using a richer and more diverse dataset can minimize the aforementioned effect.

6.2 Future Work

The first aim of further work is the creation of a solid evaluation process which will not only assess the framework created but also lead to the development of ecological conclusions. The corpus used in this study can be enriched with even older texts, or texts of travellers written in other languages (e.g., German or Greek), or description of other areas for deeper temporal or spatial comparison, even with more recent texts. As the perspective of the authors of these texts play a very important role in understanding, a thorough sentiment analysis could be insightful. The NLP analysis can also be connected to an in-depth spatial analysis with the use of satellite images or historical aerial photographs for land cover change study.

7. Acknowledgments

This work was supported by the Project “D-AI-LECT – Digital Analysis and Recognition of Handwritten Documents of Greek Dialects” (Archimedes, Athena Research Centre, Greece) and the GREEN TALENT Horizon Europe project (Grant Agreement No. 101217375). GREEN TALENT is funded by the European Union’s Horizon Europe research and innovation programme; however, the views expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the granting authority can be held responsible for them. The authors also thank the anonymous reviewers for their valuable comments.

8. Bibliographical References

- Abdelmageed, N., Löffler, F., Feddou, L., Algergawy, A., Samuel, S., Gaikwad, J., and König-Ries, B. (2022). BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10:e89481.
- Avcioğlu, A., Demir, O., and Görüm, T. (2025). An automated approach for developing geohazard inventories using news: integrating natural language processing (NLP), machine learning, and mapping. *Natural Hazards and Earth System Sciences*, 25:2421–2435.
- Barz, C., Siegel, M., and Hanss, D. (2025). Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 68–76, Tallin, Estonia, March.
- Brando, C. and Frontini, F. (2017). Semantic historical gazetteers and related NLP and corpus linguistics applications. *Journal of Map & Geography Libraries*, 13(1):1–6.
- D’Souza, J., Laubach, Z., Al Mustafa, T., Zarriß, S., Frühstückl, R., and Illari, P. (2025). Mining for species, locations, habitats, and ecosystems from scientific papers in invasion biology: A large-scale exploratory study with large language models. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 16–23, Tallin, Estonia, March.
- Grasso, F. and Locci, S. (2024). Assessing generative language models in classification tasks: Performance and self-evaluation capabilities in the environmental and climate change domain. In *International Conference on Applications of Natural Language to Information Systems*, pages 302–313, Cham, Switzerland, June. Springer Nature Switzerland.
- Gregory, I., Donaldson, C., Murrieta-Flores, P., and Rayson, P. (2015). Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9(1):1–14.
- Grossner, K., Janowicz, K., and Keßler, C. (2016). Place, period, and setting for linked data gazetteers. In *Placing Names: Enriching and Integrating Gazetteers*, pages 80–96, Indiana University Press, United States.
- Haider, T., Perschl, T., and Rehbein, M. (2025). Quantification of biodiversity from historical survey text with LLM-based best-worst-scaling. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology 2025)*, pages 61–67, Tallin, Estonia, March.
- Havinga, I., Marcos, D., Bogaart, P., Tuia, D., and Hein, L. (2024). Understanding the sentiment associated with cultural ecosystem services using images and text from social media. *Ecosystem Services*, 65:101581.
- Horde, P. and Purcell, N. (2000). *The Corrupting Sea: A Study of Mediterranean History*. Blackwell Publishing, Oxford, UK.
- Huai, S. and Van de Voorde, T. (2022). Which environmental features contribute to positive and negative perceptions of urban parks? A cross-cultural comparison using online reviews and Natural Language Processing methods. *Landscape and Urban Planning*, 218:104307.
- Kolovos, E. (2017). “There was a garden...” *The Economy of the Mediterranean Island of Andros according to the Ottoman Land and Property Survey of 1670*. Crete University Press, Herakleio, Greece.
- Kong, I., Sarmiento, F. O., and Mu, L. (2023). Crowdsourced text analysis to characterize the US National Parks based on cultural ecosystem

- services. *Landscape and Urban Planning*, 233:104692.
- Lai, K., Porter, J. R., Amodeo, M., Miller, D., Marston, M., and Armal, S. (2022). A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. *Information Processing & Management*, 59(1):102735.
- Li, T. and Shi, T. (2025). High-resolution climate reconstruction from historical Chinese weather records using optimized natural language processing. *Scientific Reports*, 15:44447.
- Lin, H. and Safi, T. (2024). ollamar: An R package for running large language models. *Journal of Open Source Software*, 10(105): 7211.
- Luo, H., Zhang, Z., Zhu, Q., Ameer, N. E. H. B., Liu, X., Ding, F., and Cai, Y. (2025). Using large language models to investigate cultural ecosystem services perceptions: A few-shot and prompt method. *Landscape and Urban Planning*, 258:105323.
- Nundloll, V., Smail, R., Stevens, C., and Blair, G. (2022). Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. *Heliyon*, 8(10).
- Schimanski, T., Senni, C. C., Gostlow, G., Ni, J., Yu, T., and Leippold, M. (2023). Exploring nature: Datasets and models for analyzing nature-related disclosures. *arXiv preprint arXiv:2312.17337*.
- Silge, J. and Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3): 37.
- Sodoge, J., Kuhlicke, C., and de Brito, M. M. (2023). Automated spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, 41:100574.
- Tomscha, S. A., Sutherland, I. J., Renard, D., Gergel, S. E., Rhemtulla, J. M., Bennett, E. M., and Clark, E. E. (2016). A guide to historical data sets for reconstructing ecosystem service change over time. *BioScience*, 66(9):747–762.
- Villacampa-Porta, J., Coronado-Vaca, M., and Garrido-Merchán, E. C. (2025). Impact of EU non-financial reporting regulation on Spanish companies' environmental disclosure: a cutting-edge natural language processing approach. *Environmental Sciences Europe*, 37(1):29.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY.
- Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., and Bryan, M. J. (2019a). Package 'readxl'. Version 1.3.1.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019b). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.
- Wijffels, J., Straka, M., and Straková, J. (2018). Package 'udpipe'. CRAN.
- Zhang, Z., He, Z., Su, K., Wu, S., and Liu, L. (2026). Global evolution of ecosystem services research in watersheds: Insights from large language models. *Land Use Policy*, 160:107849.
- Zhu, J. J., Jiang, J., Yang, M., and Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology*, 57(46):17667–17670.

Retrieving Floods without Floodlights: Topic Models as Binary Classifiers for Extreme Climate Events in German News

Brielen Madureira^{1,2}, Mariana Madruga de Brito², Andreas Niekler^{1,3}

¹LeipzigLab - Climate Discourse, Leipzig University, Germany

²Helmholtz Centre for Environmental Research - UFZ, Germany

³ Computational Humanities, Leipzig University, Germany

brielen.madureira@uni-leipzig.de

mariana.brito@ufz.de

aniekler@informatik.uni-leipzig.de

Abstract

In studies of media coverage of extreme climate events, NLP methods have become indispensable for identifying relevant texts in large news databases. Still, enough annotated data to train accurate deep learning-based classifiers from scratch is often not available. Topic Models have the advantage of being both unsupervised and interpretable, but are typically used only for exploratory analysis or data characterisation. In this study, we investigate how to employ Topic Models as binary classifiers for refining the retrieval of relevant news about seven types of extreme climate events in the German media. Our method relies on the *posterior* distributions estimated by Topic Models to select relevant documents, without modifying their training procedure. Using an annotated sample to guide the evaluation, we show that the probabilities assigned to keywords used to query news databases can also be informative for selecting relevant topics and improve sample precision. We compare our results to a fine-tuned text embedding classifier and an open-weight LLM, discussing observed trade-offs, e.g. the LLM's lowest precision. Moreover, we show that results are hazard-dependent, which speaks against considering climate events as a single category in NLP tasks.

Keywords: extreme climate events, German news, topic models, text classification, document retrieval

1. Introduction

Assume we are gathering news about floods events to study collective attention in the media. Simply querying a news database to retrieve documents containing the string *flood* would not only match news reporting on actual floods, but also many false positives. Consider this (obviously constructed) example: “*Soccer fans experienced a flood of emotions witnessing floodlights being turned on as players flooded the field: the game could finally begin after the risk of a flash flood was ruled out.*” This illustrates a central challenge in information retrieval: the term *flood* can have metaphorical senses, be part of compound nouns unrelated to climate or refer to a merely hypothetical hazard. Thus, despite the repeated presence of the term *flood*, this text is rendered unrelated to actual flood events.

Pitfalls like that can emerge at the intersection of environmental and social sciences, such as in text-based climate impact and adaptation research. This field often relies on NLP methods to process texts about climate events and their consequences (Alencar et al., 2024; Nunes Carvalho et al., 2024, *inter alia*). In this context, dictionary-based retrieval is a typical procedure: large databases are queried using a curated list of hazard-related keywords to find potentially relevant documents about e.g. floods, droughts or wildfires (e.g. Sodge et al.,

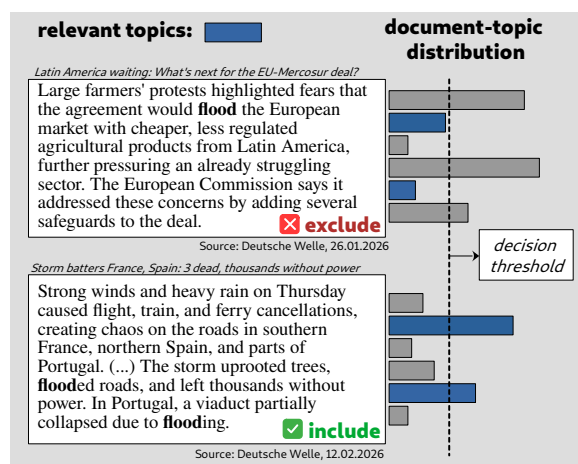


Figure 1: Relevant news articles can be identified based on the saliency of relevant topics in their representation estimated by a topic model.

2023; Li et al., 2025). But as we just saw above, the mere presence of a keyword in a document does not guarantee its relevance. If term presence or frequency are directly used as predictors in quantitative assessments, research validity is impaired.

Although keyword-based matching serves as a reasonable prefiltering step for creating an initial collection of documents with good recall, retrieval results must be further refined using other classi-

fication methods to detect true and false positives and improve their precision. This is an open problem recently discussed by Grasso et al. (2024).

In this paper, we investigate the possibility of yielding a binary classification model for identifying relevant documents using the probabilities estimated by unsupervised vanilla Topic Models (TM), as the overview in Figure 1. We assume a situation with a small amount of annotated data that is not enough for training deep learning-based models from scratch but still informative for evaluation.

Our main contributions in this paper are:

- i a data analysis of news articles in German annotated with seven types of extreme climate events;
- ii the usage of TMs for relevance classification without any needed modification on the training regime and no direct human effort in topic interpretation; and
- iii evidence that TMs are, for some hazards, on par with deep learning alternatives, with the advantage of interpretability and a tendency to higher precision.

2. Related Literature

Retrieval of environment-related documents

Document retrieval is an ubiquitous step in creating corpora for socio-environmental research. To name a few recent large-scale approaches, Leippold and Varini (2020) implemented a graph-based heuristic on Wikipedia metadata of entries on climate topics, Kong and Purves (2026) relied on climate-related keywords to retrieve news and Cai et al. (2025) used a hazard event database for a targeted query of news articles and refined results using a Large Language Model (LLM). Our work focuses on the step of *refining* an initial sample of documents retrieved via keyword-matching methods.

Topic Models TMs such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) yield distributions of topics in documents in an unsupervised fashion. Many variations exist, e.g. keyword-assisted TMs (Eshima et al., 2024), which exploit keywords to guide clustering and circumvent post hoc topic interpretation, CorEx (Gallagher et al., 2017), which relies on an information-theoretic framework, and Top2Vec (Angelov and Inkpen, 2024), which performs clustering in an embedding space shared by documents and words. TMs often aid characterising corpora in climate and ecology research (Lesnikowski et al., 2019; Stede et al., 2023; Madruga de Brito et al., 2025; Zander et al., 2023; Peura et al., 2025; Beckles and Heidke, 2025; Barz et al., 2025, *inter alia*).

Tuning and assessing the quality of TMs is intricate if performed purely intrinsically (Maier et al., 2021), but our evaluation is enhanced by annotated data that allows known document properties to be compared to the formed topics.

TMs for text classification TMs have been widely used to map texts to classes, e.g. by feeding their outputs as input features for other classifiers (Li et al., 2016b; Anantharaman et al., 2019; Seifollahi et al., 2021). Other works aligned topics to classes, either directly (Sarioglu et al., 2013), by experts (Hingmire et al., 2013) or by configuring the *priors* in a way that induces desired clusters (Miller et al., 2016; Rubin et al., 2012), e.g. by relying on relevant keywords (Chen et al., 2015; Zha and Li, 2019; Li et al., 2016a, 2018). Keyword selection can also derive a lower dimensional set of features for other types of classifier models (Onan et al., 2016). McAuliffe and Blei (2007) incorporated a response variable into the TM training, to jointly model documents and their classes or scores. While many procedures require adjusting *priors* or the modelling approach, we stick to standard LDA and NMF implementations, which are arguably more accessible for newcomers and researchers from other fields.

Text classification in climate research Climate-related text classification is an established NLP task; in many settings, it remains an unsolved problem even for LLMs, with performance often well below 0.75 F1 in the ClimateEval benchmark (Kurfali et al., 2025). In the study by Li et al. (2024), a fine-tuned encoder achieved an F1 of 0.98 for identifying relevant documents on climate extreme impacts, but only in English and on a small sample of cleaner Wikipedia entries with climate-related keywords *in their titles*. This restriction likely ensured a majority of relevant matches, but resulted in an unknown number of missed cases. When full texts are considered (as we do), there is less room for false negatives while substantially increasing the need for filtering out false positives, especially with imbalanced datasets.

The problem we tackle in this paper is similar to the work by Grasso et al. (2024): corpus construction via keyword-based prefiltering and automatic classification. We differ by focusing on German, handling specific hazards separately and exploring TMs for classification, not only for topic analysis as that work did. Our design builds upon existing work with a novel perspective: we do not change the LDA and NMF internal mechanisms and explore the *posterior* probabilities (or normalised scores) they assign to keywords as a means to automatically partition topics and perform binary classification of news about extreme climate events.

3. Methods

This section formalises the task and explains how topic models are applied for binary classification. Then, it describes the two deep learning strategies used for comparison.

3.1. Task Formalisation

Let D be a set of documents d , each belonging to a binary class $C = \{0, 1\}$, and V be the set of all tokens w that appear in D . Class 1 represents relevant documents. A document classifier is a function $f: D \rightarrow C$ that maps documents to classes and can be approximated by various methods.

Furthermore, let $F \subseteq V$ be a set of feature tokens selected from V based on given criteria (e.g. minimum frequency and part-of-speech tags) and $K \subset F$ be a small set of predefined tokens of interest which we name *keywords*. A trained topic model M with n topics T estimates two distributions: p_{feat} , the probability of a feature token in a topic and p_{topic} , the probability of a topic in a document. In other words, a document is represented as a probability distribution over topics and a topic, as a probability distribution over feature tokens.

With M 's estimations, we can define a binary relation R between D and T representing whether each document is related to each topic. To use M for binary classification, a partition of topics T with two sets is created, each corresponding to a class in C . The class of a topic must also correctly classify documents related to such a topic.

3.2. Classification with Topic Models

Firstly, a TM is trained on the entire collection of unique documents using selected hyperparameters, following standard procedures (described in Section 5). Then, two further steps are needed: (i) assigning topics to documents and (ii) identifying which topics are to be regarded as relevant.

For (i), we define the relation R as $p_{topic}(t, d) \geq \theta$ with $0 \leq \theta \leq 1$. That means that if the proportion of a topic t in a document d is at least a threshold, we consider that d discusses topic t (as in Figure 1).

For (ii), we propose two ways to partition topics into two classes, relevant and not relevant, avoiding the usual *post hoc* human interpretation in TMs:

- **keyword proximity:** topic t is assigned to the relevant class if $\exists w \in K : p_{feat}(w, t) > \gamma$. In other words, if the topic assigns a high enough probability to at least one keyword, the topic belongs to the partition of the relevant class.
- **top terms:** if there is a keyword among the top k features of a topic (ranked by probability), the topic is assigned to the relevant class.

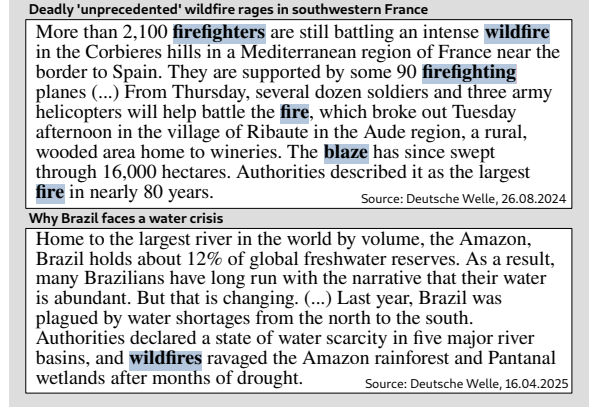


Figure 2: Wildfires as the main topic (top) or mention (bottom) in news excerpts.

The actual classification of each document is made as follows: if the document is related to at least one of the relevant topics, we consider it to be relevant. Otherwise, it is classified as not relevant.

3.3. Deep Learning Classifiers

The performance of our TM approach is compared to two deep learning alternatives: a fine-tuned text embedding model and an LLM. The first is a binary classifier trained using the SetFit framework (Tunstall et al., 2022) which fine-tunes a pretrained text embedding model with a classification head, aiming at optimising task-specific embeddings based on a set of contrastive examples. The latter prompts an LLM to generate a binary label classifying the document as relevant or not. The implementation details are explained in Section 5.

4. Data

The data for this study derives from an ongoing project on the collective attention to extreme climate events in the German media. Seven types of hazards were selected (cold waves, droughts, floods, heat waves, landslides, storms and wildfires). The wiso-net news aggregation database¹ was queried using a pre-defined list of hazard-related keywords, similar to (Li et al., 2024; Madrugada de Brito et al., 2025; Carvalho et al., 2025, see Appendix). The retrieved collection contained 13,771,411 German news articles from around 370 outlets, spanning from 2000 to 2024, split into separate sub-collections for each extreme climate event.

We make a distinction between two types of relevant news: *main*, in which the extreme climate event is the main topic, and *mention*, that refer to it *en passant*, of secondary importance among other more prominent topics, as shown in Figure 2. Both

¹<https://www.wiso-net.de/>

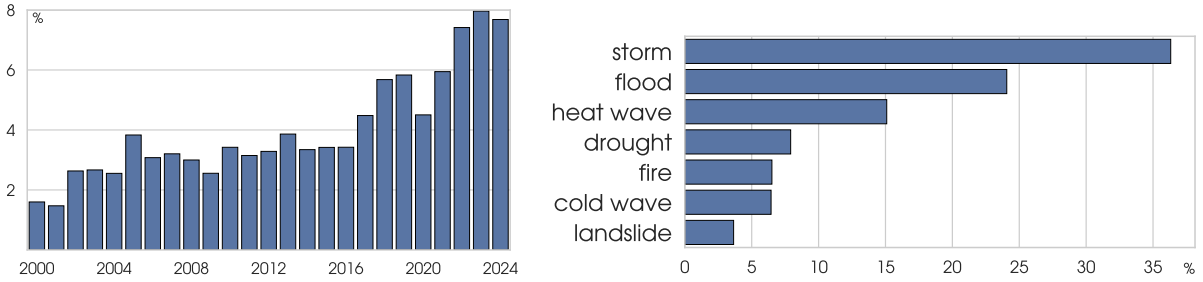


Figure 3: % of documents per year (left) and % of documents per extreme climate event (right).

forms count towards levels of collective attention, but automated identification of the latter is more challenging due to brevity and underspecification.

We are primarily interested in the coverage of *international* events in German news, so basic rule-based procedures were implemented to reduce the amount of local or unrelated news. As the data we are working with was queried via unrestricted keyword matching, many false positives occurred. Of particular relevance to this paper is the filter for what we call “intruder” keywords, i.e. words that derive from valid hazard-related keywords but are unrelated to climate events. For instance, the name *Dürrenmatt*, retrieved by *Dürre* (drought), *Flutlicht* (floodlight), by *Flut* (flood) and *Stürmer* (forward player in football), by *Stürme* (storms). To reduce the number of unrelated documents, we excluded all instances that contained only intruder keywords. Other preprocessing steps for filtering and cleaning the data are summarised in the Appendix.

Final document collection The previous steps resulted in a sample of 2,438,275 documents (17.71% of the originally retrieved instances). They have, on average, 537.19 tokens (std=362.06) and 28.66 sentences (std=20.12). The news database has an inherent temporal bias towards recent years, illustrated in Figure 3 (left). The distribution over types of hazards is also shown in Figure 3 (right).

Gold standard A sample of 3,150 documents was randomly selected while ensuring uniform distribution across hazard types (450 each) and years (18 each per hazard) and no duplicates per hazard. Two annotators classified the news as relevant or not (see annotation instructions in the Appendix) while also judging whether the event of interest was the news’ main topic or just a mention. Table 1 shows the percentage of relevant documents identified in the annotated sample. The initial effectiveness of the dictionary-based approach, together with the rule-based preprocessing, depends on the type of extreme climate event: while the landslide portion already reached a precision of almost 0.6, cold and heat waves stayed below 0.2.

	relevant	main topic
cold wave	14.44	4.22
drought	36.00	4.22
flood	43.33	12.44
heat wave	19.78	2.44
landslide	58.67	17.56
storm	27.56	7.33
wildfire	41.56	21.11

Table 1: Percentage of relevant documents for each type of extreme climate event in the gold standard.

Identifying relevant documents is not as straightforward as it may sound. A subset of 100 news was annotated by both annotators independently. The overall agreement proportion in the primary binary decision (relevant or not) was 0.77 ($\kappa = 0.53$), indicating that there are sources of legitimate disagreement in this decision. Apart from potential errors, disagreements may stem from differences in perception of what constitutes an *extreme* and *concrete* natural event. Some cases which may have involved such subjectivity were (translated from German):

- **cold wave:** “...the son, travelled on Monday during the snow chaos from Cologne to Wismar in order stay by his mother.”
- **drought:** “the mine was found two weeks ago due to the low water level in the Rhine river.”
- **landslide:** “we knew the situation when a country closed the border or a street was blocked for a week due to a landslide or something else.”

5. Experiments

The varying estimated proportions of relevant documents for each hazard sample suggest that these phenomena manifest differently not only in their nature but also in their coverage and linguistic features. Therefore, each classification strategy was

conducted for each type of extreme climate event separately. The annotated sample was randomly split into a training and a test set with 350 and 100 instances, respectively, for each hazard. The presented results were computed in the test split.²

Topic models Documents were preprocessed to extract their features partially based on the procedure by Grasso et al. (2024). We used Spacy’s³ model `de_core_news_lg` to tokenize, lemmatise and label tokens with their part-of-speech tags. Tokens with less than 3 characters and stopwords were removed, as well as non-alphabetical characters. All tokens were lowercased. The feature selection involved two criteria: the term’s document frequency and part-of-speech tag. All keywords were kept as features, even if they did not meet the minimum frequency threshold, to ensure they had a chance to contribute to forming a topic. To avoid the induction of topics based on duplicated news, only one instance of texts with high similarity was included. Gensim’s⁴ implementation of the LDA and NMF methods was used to train topic models. The number of topics was a hyperparameter. For LDA, the `eta` and `alpha` arguments were set to `auto`. We run various combinations of the three hyperparameters (minimum document frequency, part-of-speech tags and number of topics) and, for each model, we computed results varying the values for k , γ , for top term and keyword proximity, and θ . For each hazard, we selected the best-performing models in the training split. Specific parameters and the final configuration that produced the results are in the Appendix. The code is available at <https://codeberg.org/briemadu/tm-as-classifier>.

Text embeddings This classifier was trained via the Small-Text (Schröder et al., 2023) wrapper implementation around SetFit (with its default configuration in HuggingFace) and Sentence Transformers (Reimers and Gurevych, 2019). In this method, the classification is performed by a logistic regression component on top of the fine-tuned text embeddings. We opted for the BAAI/bge-m3 text embeddings released by (Chen et al., 2024) due to the model’s multilingual capabilities and longer

²Note that the use of train/test splits depends on the classifier. TMs’s unsupervised fitting included all unique documents, since the objective here is not to generalise to unseen data but to optimise for topics that best fit our own documents. Still, only the train split was used to select the best model configuration to avoid overfitting to the test data in this choice. The text embeddings model used the train split for fine-tuning. The LLM was directly prompted with the test data in a zero-shot approach.

³<https://spacy.io/>

⁴<https://radimrehurek.com/gensim/>

context length (8,192 tokens), since standard Sentence Transformers that typically allow only up to 512 tokens would not suffice for longer news articles. Training was performed with a batch size of 16 instances and a learning rate of 10^{-5} .

LLM Since the purpose of this paper is not to benchmark LLM performance, we chose only one model to serve as a reference. Results were produced by `mistralai/ministral-3-14b-reasoning`.⁵ We selected an open-weight model that could be run locally and keep the data in our own infrastructure.⁶ The prompt contained instructions similar to those given to the annotators, including the definition of the hazard and of the labels, the hazard’s keywords and the main body of the news article. The exact prompt and values are in the Appendix. We had to programatically parse answers that included spurious prefixes before the actual label.

Evaluation

The models’ performance was quantitatively assessed with conventional binary classification metrics: precision, recall and F1 score of the positive class. The test sample’s precision and a presumed recall of 1 were used as a baseline to measure how much the classifiers improve retrieval precision without reducing its recall. The evaluation was enriched with a detailed analysis of the TM results.

We present results for three variations of TMs: `TM-F1` was run with the configuration that resulted in the highest F1 score (on the training split) in our hyperparameter search; `TM-B` uses the configuration that balanced precision and recall to be both as high as possible; and `TM-P` has the configuration with the highest precision while retaining some level of recall. We also compare results to an ensemble strategy that performs classification via majority voting across the outputs of `TM-B`, fine-tuned text embeddings and LLM classifiers.

6. Results

Aggregated results We first examine results aggregated over the whole test split ($n = 700$), i.e. including all extreme climate events. Table 2 shows precision, recall and F1 score for all classifiers. The rightmost column shows the number of news articles of type `main` that were correctly identified as relevant. All classifiers succeeded in considerably increasing the low proportion of relevant documents in the keyword-based sample, but TMs and deep

⁵<https://huggingface.co/mistralai/Ministral-3-14B-Reasoning-2512>

⁶We did not compare results to closed commercial models as they are at odds with open science principles.

	P	R	F1	n_{main}
baseline	0.350	1.000	0.519	58
TM-F1	0.637	0.710	0.672	56
TM-B	0.710	0.649	0.678	55
TM-P	0.808	0.396	0.532	47
fine-tuned	0.647	0.853	0.736	57
llm	0.583	0.976	0.730	58
majority	0.701	0.890	0.784	58

Table 2: Aggregated results: binary precision, recall and F1 score of all classifiers in the test split and the number of news of type `main` correctly identified.

learning strategies behaved differently in how precision and recall were balanced. While the LLM had almost maximum recall with a substantial margin over other models, its precision was the lowest. `TM-P` had the highest precision but at the cost of low recall. `TM-B` achieved the second highest precision with a more reasonable recall. The majority voting approach led to the highest F1 score. If we focus on the identification of news of type `main`, all classifiers (apart from `TM-P`) performed very well, identifying at least 55 out of the 58 instances.

Results by hazard Aggregated results can mask variations in performance for each underlying hazard. Table 3 summarises results by hazard type, in line with the fact that models were trained separately. We can see that metrics varied greatly depending on the phenomenon: the lowest best F1 score of 0.59 occurred for heat wave whereas the highest best of 0.92 was observed for landslide. Majority voting achieved the best F1 scores for five hazards and the fine-tuned text embeddings for the other two. The LLM consistently held the highest recall in all hazards. `TM-B` had the best precision in the three most imbalanced (cold waves, heat waves and storms).

Discussion In aggregated results, TM performance was indeed lower than that of deep learning strategies, but the moderate reduction of only around 0.06 in F1 score still provides a much-desired gain in interpretability: we can explain exactly why each document was classified as relevant. The deep learning strategies tended to incur more false positives whereas TMs could reduce the proportion of unrelated documents while causing more false negatives. Models with higher precision but low recall, like `TM-P`, can still be useful when sample precision is a priority, since a sample with low recall may still be representative and of enough size in large datasets. High precision helps reduce the impact of unrelated documents in downstream anal-

		P	R	F1
cold wave	baseline	0.170	1.000	0.291
	TM-F1	0.471	0.471	0.471
	TM-B	0.583	0.412	0.483
	TM-P	0.500	0.059	0.105
	fine-tuned	0.297	0.647	0.407
	llm	0.455	0.882	0.600
	majority	0.542	0.765	0.634
drought	baseline	0.440	1.000	0.611
	TM-F1	0.517	0.682	0.588
	TM-B	0.622	0.523	0.568
	TM-P	0.938	0.341	0.500
	fine-tuned	0.686	0.795	0.737
	llm	0.525	0.955	0.677
	majority	0.692	0.818	0.750
flood	baseline	0.360	1.000	0.529
	TM-F1	0.605	0.639	0.622
	TM-B	0.595	0.611	0.603
	TM-P	0.750	0.167	0.273
	fine-tuned	0.737	0.778	0.757
	llm	0.600	1.000	0.750
	majority	0.738	0.861	0.795
heat wave	baseline	0.200	1.000	0.333
	TM-F1	0.423	0.550	0.478
	TM-B	0.600	0.450	0.514
	TM-P	0.600	0.150	0.240
	fine-tuned	0.439	0.900	0.590
	llm	0.322	0.950	0.481
	majority	0.429	0.900	0.581
landslide	baseline	0.580	1.000	0.734
	TM-F1	0.785	0.879	0.829
	TM-B	0.778	0.845	0.810
	TM-P	0.816	0.690	0.748
	fine-tuned	0.877	0.983	0.927
	llm	0.826	0.983	0.898
	majority	0.826	0.983	0.898
storm	baseline	0.270	1.000	0.425
	TM-F1	0.680	0.630	0.654
	TM-B	0.800	0.593	0.681
	TM-P	0.625	0.185	0.286
	fine-tuned	0.558	0.889	0.686
	llm	0.614	1.000	0.761
	majority	0.714	0.926	0.806
wildfire	baseline	0.430	1.000	0.601
	TM-F1	0.773	0.791	0.782
	TM-B	0.825	0.767	0.795
	TM-P	0.844	0.628	0.720
	fine-tuned	0.750	0.837	0.791
	llm	0.662	1.000	0.796
	majority	0.809	0.884	0.844

Table 3: Detailed results: binary precision, recall and F1 score of all classifiers in the test split for each hazard type.

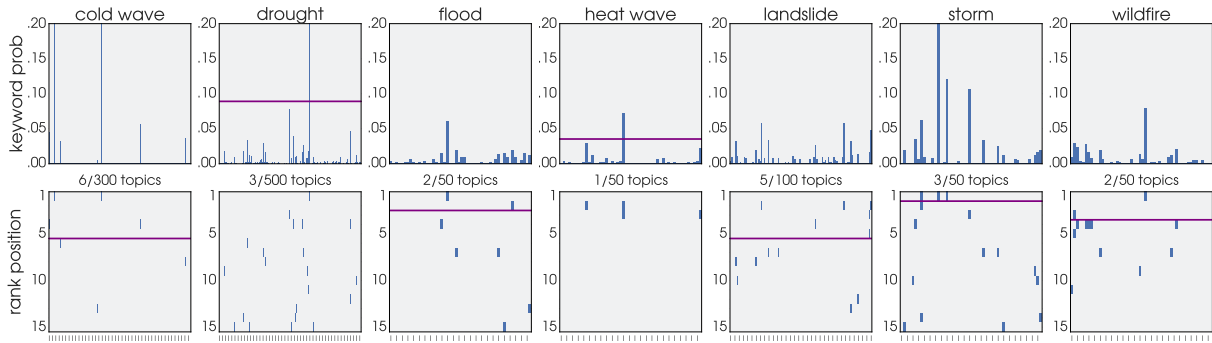


Figure 4: High-level overview of how relevant topics were identified by TM-B . In the first row, bars represent the maximum probability each topic (x-axis) assigns to keywords. The second row highlights in dark the positions in the rank that contain keywords in each topic (x-axis). The purple horizontal bars represent the optimal γ and k , respectively, in our experiments (see Appendix for exact numerical values).

yses. For situations in which news of type `main` are more important than `mention`, the two best TMs and deep learning classifiers worked in like manner, not missing the vast majority of instances.

The majority voting approach seemed to draw out advantages from each model, achieving the best F1 score. Still, employing three computationally costly models for this task is hardly justifiable in practice, given the modest overall increase in the aggregated F1 score compared to single models.

Classification of news turned out to be hazard-dependant. There was no one-size-fits-all best solution across all hazards. The fine-tuned text embeddings balanced precision and recall well in general, but in three hazards (cold waves, storms and wildfires) the F1 score of the TM approaches was on par or better than it, which is a very interesting finding given that TMs are unsupervised and do not rely on the currently prevailing deep learning paradigm. Landslides and wildfires were the easiest to identify with all metrics above 0.75 (except for the LLM’s precision for wildfire) in all models apart from TM-P . Cold and heat waves were the most challenging with suboptimal results even for the majority voting method.

Note, however, that comparisons between models should be done with caution, as these experimental estimates by hazard type were computed from samples of only 100 documents each. Rare events become very sensitive to individual predictions in such a small sample. For instance, cold waves contain only 17 relevant documents on which to measure precision and recall, so that a single swapped prediction by a model would already cause a 5.8% increase or decrease in recall.

7. Analysis

In this section, we explore TM’s interpretability by providing more details on the TM-B models’ behaviour. In our non-exhaustive hyperparameter

DROUGHT			WILDFIRE	
dürre	mitte	trockenheit	feuer	buschfeuer
notstand	niedrigwasser	waldbrandgefahr	brand	australisch
ernteausfall	tag	feuchtigkeit	waldbrand	koala
vieh	fisch	kanton	flamme	bundesstaat
regenzeit	wasser	brandgefahr	waldbrand	kontinent
ausmaß	stoff	ernteausfall	hektar	buschbrand
zentrum	fischsterben	nässe	feuerwehr	australier
versicherung	sand	stress	region	tier
wasserreservoir	kreis	notstand	groß	villa
helmholtz	jugendliche	leiter	kontrolle	ostküste

Figure 5: Top 10 terms in each of the topics considered as relevant for drought and wildfire.

search, LDA achieved the best performance for six hazard types, while NMF was superior only for drought. The optimal thresholds θ for assigning topics to documents were between 0.028 and 0.076. Figure 4 illustrates how relevant topics were selected. The top terms decision method achieved the highest performance across five hazards with k values ranging from 1 to 5. Keyword proximity was superior only for drought and heat wave using $\gamma = 0.09$ and 0.036, respectively. The number of selected topics for each hazard varied from 1 to 6.

Here we focus on wildfires and droughts as they had the smallest and largest differences in F1 score, respectively, in relation to deep learning strategies. For wildfires, topics were considered relevant if a hazard-related keyword was among their top 3 most probable terms. That resulted in 2 out of 50 topics being considered as relevant. For drought, keyword proximity selected 3 out of 500 topics as relevant. The top 10 lemmas representing these topics are shown in Figure 5. Figure 6 illustrates the effect of θ for wildfire’s leftmost relevant topic: how well (not) relevant documents are classified based on the θ parameter for the rightmost relevant topic in fire: documents with topic probability above θ are classified as relevant, with a few wrong predictions.

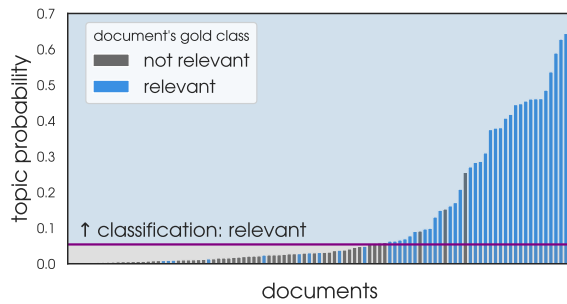


Figure 6: Example of the θ threshold for assigning a relevant fire topic to documents. Above the purple horizontal line, documents are classified as relevant, with a few mistakes with respect to the gold standard.

Wildfires Out of the 100 test instances, 83 were correctly classified. We inspected the 17 errors. The 7 false positives are texts that do refer to fires, but they are either not wildfires (e.g. fires in houses or industrial plants) or discuss technologies to combat wildfires. One of the documents describes a wildfire warning that also refers to an extinguished fire, which may have been missed by the annotators. Among the 10 false negatives, most contain mentions to wildfires occurring in discussion about other main topics (financial markets, heat waves, conferences) in documents that are a concatenation of various news articles (a problem we inherited from the original database were not able to fully solve automatically). In such cases, the fire-related topic may not have been salient enough to be assigned to the document. Although two topics were selected as relevant, only the left one was responsible for making all predictions on the test set. The second topic appeared in only one text, indicating that, although potentially relevant, it may have specialised too much during training. All documents of type `main` were correctly identified.

Drought For this hazard, 65 documents were correctly classified, with 14 false positives and 21 false negatives. Many false positives treat drought as a broad phenomenon rather than a concrete event, for example, when discussing drought-tolerant plants, vegetation stress, or climate change. Such cases are difficult to distinguish using TMs and may also reflect ambiguities in the annotation. False negatives show no clear patterns. Since this model relied on 500 topics, the drought concept may have been diffuse across multiple topics: in Figure 4, it is evident that drought keywords appear in various topics not selected as relevant. The only `main` document that was incorrectly classified has only one keyword, a compound noun (*Dürregebieten*), which was not included as a feature. In-

cluding all words *containing* keywords as features could have prevented this but it introduces additional noise from low-frequency terms that form topics.

8. General Discussion

This work was primarily motivated by the lack of a comprehensive global database of extreme climate disasters. Existing disaster databases, for instance the EM-DAT (Delforge et al., 2025), are shaped by reporting practices and inclusion thresholds (e.g. at least 10 fatalities), which have been widely discussed for their biased coverage toward large, well-documented events and wealthier regions, systematically under-representing some regions and hazards (Jones et al., 2023). Our method is designed to support bottom-up data-driven analyses by bypassing the inherent incompleteness and structural biases of top-down lists of worldwide extreme climate events (Gall et al., 2009). Our procedure permits the inclusion of news about events that did not meet the arbitrary inclusion criteria of disaster databases.

Rather than claiming the superiority of one model type for news classification, we have provided evidence that the results are hazard-dependent. This is an important finding for climate impact and adaptation research: the way different hazards are reported in the news varies, so solutions that treat all climate-related hazards as a single category (i.e. disasters in general) risk masking important performance variations, leading subsequent conclusions to be biased towards those that are easier to identify. In this context, an advantage of our approach is that we explicitly consider hazards separately, enabling more reliable downstream analyses.

The exact reasons for such differences require further investigation. First, each hazard is inherently distinct in the abruptness of its onset, its duration, its frequency, and its perceived severity. Then, media coverage can differ depending on socioeconomic and geo-political factors. Finally, there is linguistic and discourse-related variation. For instance, while some keywords are very specific to climate events, others are polysemic and appear in multi-word expressions. The interdependencies among these layers are worth studying. Some events are hard to pinpoint even for humans, which can impact gold standards. Treating extreme climate events as a monolithic concept is thus not advisable in NLP tasks. Besides, since multi-hazard events occur in reality, another promising way forward is to analyse how they also co-occur in news.

We aimed to reduce human input in TM interpretation by selecting thresholds automatically and minimizing hyperparameter choices. Further work

can investigate whether manual selection of keywords and topics can improve results. Our preliminary experiments with CorEx and Top2Vec yielded comparable results, so we prioritised the more traditional LDA/NMF methods in this study. However, other TM variations can be further investigated, including tuning priors to promote clearer keyword-related topics. The fine-tuned text embeddings achieved some of the highest F1 scores using only 350 documents and can potentially be further improved with active learning (Schröder and Niekler, 2020).

LLMs are being uncritically employed for many NLP tasks. We have shown that even a model with 14b parameters was not sufficiently precise. Our results add to the evidence that LLMs require careful evaluation as any other model. If LLM-based approaches are to be used, TMs can still be helpful in shrinking the amount of unrelated documents (e.g. by excluding those that have high probability for totally unrelated topics), thereby reducing the considerable environmental and financial costs of using LLMs.

9. Conclusion

We have presented a comparative analysis of three binary classifiers for refining collections of news articles on extreme climate events retrieved via keyword-based approaches. Although the LLM and the fine-tuned text embeddings had a higher F1 score in general, the drop in comparison to TMs was 0.148 on the worst case (drought) but also only 0.001 on the best case (wildfire). This is remarkable given TMs' unsupervised training and the simplicity of the keyword-guided topic selection process. Depending on the use case, this difference may be acceptable given other advantages, such as higher precision. Besides, the reason for deep learning-based predictions are beyond human comprehension, whereas decisions based on TMs are fully transparent and explainable.

Limitations

The rule-based filtering may have excluded relevant documents, although it was a price worth paying to reduce the immense volume of unrelated news and to keep the task computationally tractable. Although we are seeking to identify extreme climate events, other types of disasters (e.g. urban fires and industrial accidents that cause dam collapse) could not yet be fully distinguished by our methods.

The test samples for each extreme climate event contain only 100 documents each, which may obscure variance in the estimates. More definitive claims about differences in models' behaviour require cross-validation and, ideally, a larger sample.

The performance of the classifiers is bounded by the quality of the annotation. Despite best efforts, ambiguity is not always easy to resolve and arbitrary decisions can impact models' training and evaluation.

We presented results for varying TM set-ups as we opted for selecting the best-performing configurations. Still, keeping it constant would facilitate the direct comparison across hazards. The hyperparameter search for TM considered only a few dozen combinations of the number of topics, POS-tags and minimum document frequency. This can potentially be further refined for each hazard separately.

We did not perform extensive prompt engineering for the LLM, as these models are supposed to parse natural language instructions well; still, given their unpredictable nature, minor changes to the prompt might have led to different outcomes. Larger models may yield better results, but our focus here was on lower-scale, local solutions.

Acknowledgements

We thank Marc Keuschnigg for his contribution in conceptualising the research project that motivates this paper, as well as Maike Reichel and Julius Hehenkamp for their help in annotating the data. We also thank the anonymous reviewers for their valuable feedback.

10. Bibliographical References

- Pedro Henrique Lima Alencar, Jan Sodoge, Eva Nora Paton, and Mariana Madruga De Brito. 2024. [Flash droughts and their impacts—using newspaper articles to assess the perceived consequences of rapidly emerging droughts](#). *Environmental Research Letters*, 19(7):074048.
- Aditya Anantharaman, Arpit Jadiya, Chandana Tulasai Sai Siri, Bharath NVS Adikar, and Biju Mohan. 2019. [Performance evaluation of topic modeling algorithms for text classification](#). In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 704–708.
- Dimo Angelov and Diana Inkpen. 2024. [Topic modeling: Contextual token embeddings are all you need](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA. Association for Computational Linguistics.
- Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. [Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions](#).

- In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76, Tallinn, Estonia. University of Tartu Library.
- Valentina Tretti Beckles and Adrian Vergara Heidke. 2025. [Thematic categorization on pineapple production in Costa Rica: An exploratory analysis through topic modeling](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 44–55, Tallinn, Estonia. University of Tartu Library.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. [Latent dirichlet allocation](#). *The Journal of Machine Learning Research*, 3:993–1022.
- Erica Cai, Xi Chen, Reagan Grey Keeney, Ethan Zuckerman, Brendan O’Connor, and Przemyslaw A. Grabowicz. 2025. [Identifying and investigating global news coverage of critical events such as disasters and terrorist attacks](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):307–323.
- Tais Maria Nunes Carvalho, Andreas Niekler, Christian Kuhlicke, Jakob Zscheischler, and Mariana Madruga de Brito. 2025. [Global synthesis of peer-reviewed articles reveals blind spots in climate impacts research](#). Preprint, available at Research Square.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. [Dataless text classification with descriptive lda](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).
- Damien Delforge, Valentin Wathélet, Regina Below, Cinzia Lanfredi Sofia, Margo Tonnelier, Joris A.F. van Loenhout, and Niko Speybroeck. 2025. [Em-dat: the emergency events database](#). *International Journal of Disaster Risk Reduction*, 124:105509.
- Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2024. [Keyword-assisted topic models](#). *American Journal of Political Science*, 68(2):730–750.
- Melanie Gall, Kevin A. Borden, and Susan L. Cutter. 2009. [When do losses count?: Six fallacies of natural hazards loss data](#). *Bulletin of the American Meteorological Society*, 90(6):799 – 810.
- Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. [Anchored correlation explanation: Topic modeling with minimal domain knowledge](#). *Transactions of the Association for Computational Linguistics*, 5:529–542.
- Francesca Grasso, Ronny Patz, and Manfred Stede. 2024. [NYTAC-CC: A climate change sub-corpus of New York Times articles](#). In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 403–409, Pisa, Italy. CEUR Workshop Proceedings.
- Swapnil Hingmire, Sandeep Chougule, Girish K. Palshikar, and Sutanu Chakraborti. 2013. [Document classification by topic labeling](#). SIGIR ’13, page 877–880, New York, NY, USA. Association for Computing Machinery.
- Rebecca Louise Jones, Aditi Kharb, and Sandy Tubeuf. 2023. [The untold story of missing data in disaster research: a systematic review of the empirical literature utilising the emergency events database \(em-dat\)](#). *Environmental Research Letters*, 18(10):103006.
- Inhye Kong and Ross S. Purves. 2026. [Analyzing geographic bias of newspaper articles reporting global climate disasters](#). *Annals of the American Association of Geographers*, 116(2):270–288.
- Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. 2025. [ClimateEval: A comprehensive benchmark for NLP tasks related to climate change](#). In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207, Vienna, Austria. Association for Computational Linguistics.
- Markus Leippold and Francesco Saverio Varini. 2020. [Climatext: A dataset for climate change topic detection](#). In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.
- Alexandra Lesnikowski, Ella Belfer, Emma Rodman, Julie Smith, Robbert Biesbroek, John D. Wilkerson, James D. Ford, and Lea Berrang-Ford. 2019. [Frontiers in data analytics for adaptation research: Topic modeling](#). *WIREs Climate Change*, 10(3):e576.
- Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016a. [Effective document labeling with very few seed words: A topic model approach](#). In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM ’16*, page 85–94, New York, NY, USA. Association for Computing Machinery.
- N. Li, W. Thiery, S. Zahra, M. Madruga de Brito, K. Worou, M. Kurfali, S. Lampe, P. Muñoz,

- C. Flynn, C. Trigo, J. Nivre, J. Zscheischler, and G. Messori. 2025. [Wikimpacts 1.0: A new global climate impact database based on automated information extraction from wikipedia](#). *EGUsphere*, 2025:1–43.
- Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, Gabriele Messori, and Joakim Nivre. 2024. [Using LLMs to build a database of climate extreme impacts](#). In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 93–110, Bangkok, Thailand. Association for Computational Linguistics.
- Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. 2018. [Dataless text classification: A topic modeling approach with document manifold](#). CIKM '18, page 973–982, New York, NY, USA. Association for Computing Machinery.
- Zhenzhong Li, Wenqian Shang, and Menghan Yan. 2016b. [News text classification model based on topic model](#). In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5.
- Mariana Madruga de Brito, Jan Sodoge, Heidi Kreibich, and Christian Kuhlicke. 2024. [Comprehensive assessment of flood socioeconomic impacts through text-mining](#). *Water Resources Research*, 61(1).
- Mariana Madruga de Brito, Jan Sodoge, Heidi Kreibich, and Christian Kuhlicke. 2025. [Comprehensive assessment of flood socioeconomic impacts through text-mining](#). *Water Resources Research*, 61(1):e2024WR037813.
- Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. 2021. [Applying lda topic modeling in communication research: Toward a valid and reliable methodology](#). In *Computational methods for communication science*, pages 13–38. Routledge.
- Jon Mcauliffe and David Blei. 2007. [Supervised topic models](#). In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.
- Timothy Miller, Dmitriy Dligach, and Guergana Savova. 2016. [Unsupervised document classification with informed topic models](#). In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.
- Taís Maria Nunes Carvalho, Francisco De Assis De Souza Filho, and Mariana Madruga De Brito. 2024. [Unveiling water allocation dynamics: a text analysis of 25 years of stakeholder meetings](#). *Environmental Research Letters*, 19(4):044066.
- Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. [Ensemble of keyword extraction methods and classifiers in text classification](#). *Expert Systems with Applications*, 57:232–247.
- Telma Peura, Attila Krizsán, Salla-Riikka Kuusalu, and Veronika Laippala. 2025. [Perspectives on forests and forestry in Finnish online discussions - a topic modeling approach to suomi24](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 10–15, Tallinn, Estonia. University of Tartu Library.
- 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. Association for Computational Linguistics.
- Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. [Statistical topic models for multi-label document classification](#). *Machine learning*, 88(1):157–208.
- Efsun Sarioglu, Kabir Yadav, and Hyeong-Ah Choi. 2013. [Topic modeling based classification of clinical reports](#). In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 67–73, Sofia, Bulgaria. Association for Computational Linguistics.
- Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. [Small-text: Active learning for text classification in python](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.
- Christopher Schröder and Andreas Niekler. 2020. [A survey of active learning for text classification using deep neural networks](#). ArXiv preprint: 2008.07267.
- Sattar Seifollahi, Massimo Piccardi, and Alireza Jolfaei. 2021. [An embedding-based topic model for document classification](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(3).

Jan Sodge, Christian Kuhlicke, and Mariana Madruga De Brito. 2023. [Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning](#). *Weather and Climate Extremes*, 41:100574.

Manfred Stede, Yannic Bracke, Luka Borec, Neele Charlotte Kinkel, and Maria Skeppstedt. 2023. [Framing climate change in nature and science editorials: applications of supervised and unsupervised text categorization](#). *Journal of Computational Social Science*, 6(2):485–513.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. [Efficient few-shot learning without prompts](#). ArXiv preprint: 2209.11055.

Kerstin K Zander, Hunter S Baggen, and Stephen T Garnett. 2023. [Topic modelling the mobility response to heat and drought](#). *Climatic Change*, 176(4):42.

Daochen Zha and Chenliang Li. 2019. [Multi-label dataless text classification with topic modeling](#). *Knowledge and Information Systems*, 61(1):137–160.

A. Appendix

Further details about preprocessing and filtering

The lists of German keywords used for each hazard are shown in Figures 8, 9, and 10. Although *derecho* was included as a keyword initially, texts containing only this keyword were removed in a preprocessing step. The inclusion of *Regenfälle* (rainfalls) as a keyword for flood resulted in the inclusion of some texts that may not be about floods.

Here, we provide a summary of the preprocessing and filtering steps applied to the original document collection. Exact implementation details are documented in the preprocessing code which is available upon request.

We removed exactly duplicated instances, i.e. those pairs or groups of documents for which *all* metadata values were exactly the same. Documents with the same text but published by different outlets or on different dates were kept as they count separately towards media attention.

The regex pattern '`<. *?>`' was used to remove reminiscent html content. To split (at least part of) the documents that have been concatenated as a single instance, despite being composed of several different pieces of news, we used another regex pattern with frequent news agencies abbreviations

(e.g. *dpa* and *afp*) that often appeared in parentheses in between such concatenations.

Approximately duplicated texts were identified using the MinHash algorithm to estimate Jaccard similarity, with a threshold of 0.8 slightly more conservative than the empirical choice in (Madruga de Brito et al., 2024). This was not used to exclude any document, but helped ensure that annotators did not annotate the same text (from different news sources) for the same hazard and that TMs were not trained on similar texts that would form spurious clusters.

Spacy's German model `de_core_news_lg` was used to parse each text and retrieve tokens and sentence counts.

We also applied filters to reduce the number of unrelated documents and local news. The inclusion criteria were as follows:

- The document contains at least one keyword related to its assigned hazard. Although this was an imposed criterion for the database query, after splitting concatenated documents, were a few cases of texts that no longer contained keywords.
- The document contains at least one keyword of its assigned hazard which is *not* an intruder.
- The document's outlet reportedly belongs to the German press.
- In case of exact duplicates (regarding all fields), only one instance was kept.
- The document's ressort does not contain the word `lokal`, since we are only interested in international extreme climate events.
- The number of tokens is at least 30 and no more than 1,700. The thresholds were selected based on empirical observations of the distribution's histogram and by taking into account an initial batch of annotated documents.
- The document contains at least one of the following: a country name, a nationality (as an adjective or a noun) or a city name.
- The first token of the document is not the name of a German city followed by a full stop.
- The proportion of non-alphabetical characters is less than 0.11. The threshold was selected based on empirical observations of the distributions, also considering an initial batch of annotated documents.

Further details about the annotation

Figure 7 shows the instructions given to the two annotators. They also identified the sentences that

refer to each type of hazard and the country where it occurred. These variables will be used in future studies.

Further details about the classifiers

LLM Figures 8, 9, and 10 show the values used to fill in the hazard-dependent slots in the prompt for the LLM, which is shown in Figure 11. Definitions were translated from the EM-DAT' glossary⁷, except for storm, not defined by EM-DAT, for which we used Wikipedia⁸.

Topic Models The number of iterations and passes were fixed at 400 and 20, respectively. The random seed was set to 123. Table 4 shows the number of documents used to train the TMs for each extreme climate event, i.e. the unique texts in the collection. Table 5 shows the selected hyperparameters for the topic model configuration of each extreme climate event. In Table 6 we show all topics selected as relevant for $\tau\text{M-B}$ characterised by their top 10 terms with the highest probability.

cold	91,140
drought	96,441
flood	334,607
heat	197,800
landslide	45,039
storm	488,068
fire	78,426

Table 4: Number of documents used to train the topic models for each hazard.

⁷<https://doc.emdat.be/docs/data-structure-and-content/glossary/>

⁸<https://en.wikipedia.org/wiki/Storm>

Annotation Task: Identifying news about natural hazards

You are seeing a collection of news published in the German media. We need to identify documents that are discussing **concrete, specific, natural** hazards events **abroad**. Each document has been pre-assigned to a hazard that it is possibly discussing. There are seven hazards: flood, storm, heat wave, cold wave, fire, drought and landslide/mass movement. Your task is to annotate each document with three types of information:

Decision 1: Main document label

The first step is to identify whether the document's topic is or not about the specific event it was already assigned. You will see the document and the possible category. For example, let's assume that a document is supposed to be about a flood. Pick one of the four options:

- **unrelated**: if the document is not related to any concrete natural flood event, select the option **no, the document is not about any concrete natural flood event**.
- **main**: if the main purpose of the document is to report about one or more concrete natural flood events, select the option **yes, it's the document's main topic**.
- **mention**: if there are one or more sentences mentioning concrete natural flood events, but they are not the main topic of the document (or, in other words, if it does not seem that the main intent of the journalist when writing this document was to report on those specific events), select the option **yes, but only a mention**.
- **mention-concatenated**: unfortunately, some documents in our database are made of a list of concatenated pieces of news that should have been different instances. We applied some preprocessing steps to prevent it. Still, if you identify a case where a flood is discussed in a document that is composed of various documents on different topics, select **yes, but among other concatenated documents** instead.

The word *flood* will be replaced by the corresponding hazard in each document.

Details

- We are only interested in news about specific events, i.e. events that have taken place at some location and some point in time (droughts may be harder in this regard, but still, we want documents about concrete droughts that have happened or were happening at the moment of publication).
- We are only interested in **natural** events. In particular for fires: wildfires are relevant, but e.g. a fire in an industrial plant caused by an explosion is not.
- Documents discussing impacts of natural hazards in general, the possibility or prediction of a hazard that has not yet happened, scientific studies about hazards, metaphorical uses, etc are not relevant.
- Mentions that are too generic are also not relevant for us, for instance "*Climate change has increased the occurrence of floods in Brazil*". Such cases can be classified as **unrelated**.

Decision 2: Location

If you select any of the yes-options for the first question, please also select the country where the event you identified happened (even if it is Germany). You can select more than one country if needed.

Decision 3: Type of natural hazard in text spans

Please use the hazard labels to highlight all concrete natural hazard being mentioned/reported in the text. We are interested in the following seven hazards (use **other** for any other types):

- drought, flood, storm, heatwave, coldwave, mass movement, wildfire, other

Figure 7: Instructions given to the annotators.

		model	decision method	θ	γ or k	min. doc freq	tags	topics
TM-F1	landslide	lda	top terms	0.058	5.000	100	noun, verb, adj	100
	fire	lda	top terms	0.040	3.000	50	noun, verb, adj	50
	flood	lda	top terms	0.060	2.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.024	2.000	10000	noun, verb, adj	50
	drought	nmf	keyword proximity	0.016	0.198	500	noun	500
	heat	lda	keyword proximity	0.024	0.108	100	noun, verb, propn	100
	cold	lda	top terms	0.024	5.000	500	noun, verb, adj	300
TM-B	landslide	lda	top terms	0.076	5.000	100	noun, verb, adj	100
	fire	lda	top terms	0.054	3.000	50	noun, verb, adj	50
	flood	lda	top terms	0.062	2.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.030	1.000	10000	noun, verb, adj	50
	drought	nmf	keyword proximity	0.028	0.090	500	noun	500
	heat	lda	keyword proximity	0.052	0.036	100	noun, verb, adj	50
	cold	lda	top terms	0.028	5.000	500	noun, verb, adj	300
TM-P	landslide	lda	keyword proximity	0.064	0.054	50	noun, verb, adj	100
	fire	lda	top terms	0.064	1.000	500	noun, propn	100
	flood	lda	top terms	0.054	1.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.120	5.000	10000	noun, verb, adj	50
	drought	nmf	top terms	0.034	1.000	500	noun, verb, adj	500
	heat	lda	top terms	0.148	2.000	500	noun, verb, propn	50
	cold	lda	keyword proximity	0.146	0.036	5000	noun, verb, adj	50

Table 5: Hyperparameters of all topic models that produced the presented results.

landslide :

hazard: Erdrutsch

keywords: Erdrutsch, Felssturz, Felsstürz, Schlammlawine, Massenbewegung, Hangrutsch, Hangbewegung, Rutschung, Bodenrutsch, Hangabrutschung, Murgang, Gerölllawine, Rutschhang, Rutschhäng, Rutschgefahr, Felslawine, Mure

hazard_event: Erdrutschereignisse

definition: Jede Art von mäßiger bis schneller Bodenbewegung, einschließlich Lahare, Schlammlawinen und Murgänge (unter trockenen/nassen Bedingungen). Ein Erdrutsch ist eine durch die Schwerkraft gesteuerte Bewegung von Erde oder Gestein, deren Geschwindigkeit in der Regel zwischen langsam und schnell liegt, jedoch nicht sehr langsam ist. Er kann oberflächlich oder tief sein, aber das Material muss eine Masse bilden, die einen Teil des Hangs oder den Hang selbst ausmacht. Die Bewegung muss nach unten und nach außen mit einer freien Fläche erfolgen. ODER Jede Art von Abwärtsbewegung von Erdmaterialien unter hydrologisch trockenen Bedingungen. ODER Arten von Massenbewegungen, die auftreten, wenn starker Regen oder schnelle Schnee-/Eisschmelze große Mengen an Vegetation, Schlamm oder Gestein unter dem Einfluss der Schwerkraft einen Hang hinunterbefördern.

fire :

hazard: Waldbrand

keywords: Flächenbrand, Flächenbränd, Waldbrand, Waldbränd, Wildfeuer, Landschaftsbrand, Landschaftsbränd, Buschfeuer, Vegetationsbrand, Vegetationsbränd, Naturbrand, Naturbränd, Großbrand, Großbränd, Forstbrand, Forstbränd, Heidebrand, Heidebränd

hazard_event: Waldbrandereignisse

definition: Jede unkontrollierte und nicht vorgeschriebene Verbrennung oder das Abbrennen von Pflanzen in einer natürlichen Umgebung wie Wald, Grasland, Buschland oder Tundra, die natürliche Brennstoffe verbraucht und sich aufgrund von Umweltbedingungen (z. B. Wind oder Topografie) ausbreitet. Waldbrände können durch Blitzeinschläge oder menschliches Handeln ausgelöst werden.

Figure 8: Keywords and values used in the prompts for each hazard (1/3).

```

cold :

    hazard : Kältewelle

    keywords : Kältewelle , Kälteeinbruch , Kältestress , extreme Kälte , extremer Kälte , extremen Kälte , Frost , strenger Winter , Wintereinbruch , Wintereinbruch , Kälteperiode , Kälterekord , arktische Kälte , arktischer Kälte , arktischen Kälte , Kältewarnung , Eisregen , Eiseskälte , Schneechaos

    hazard_event : Kältewellenereignisse

    definition : Eine Periode mit ungewöhnlich kaltem Wetter. In der Regel dauert eine Kältewelle zwei oder mehr Tage und kann durch starke Winde noch verstärkt werden. Die genauen Temperaturkriterien für eine Kältewelle können je nach Standort variieren .

heat :

    hazard : Hitze

    keywords : Hitze , extreme Temperaturen , extremen Temperaturen , Temperaturrekord , Tropennacht , Tropennächt , überhitzung , Rekordhitze , Hitzetag

    hazard_event : Hitzewellenereignisse

    definition : Eine Periode mit ungewöhnlich heißem und/oder ungewöhnlich feuchtem Wetter. In der Regel dauert eine Hitzewelle zwei oder mehr Tage. Die genauen Temperaturkriterien für eine Hitzewelle können je nach Standort variieren .

drought :

    hazard : Dürre

    keywords : Dürre , Rekorddürre , Trockenperiode , Trockenheit , Wasserknappheit , Niedrigwasser , Wassermangel , Niederschlagsmangel , Niederschlagsdefizit , Bodenfeuchte-Defizit , Bodenfeuchteverlust

    hazard_event : Dürreereignisse

    definition : Ein längerer Zeitraum mit ungewöhnlich geringen Niederschlägen , der zu einer Wasserknappheit für Menschen , Tiere und Pflanzen führt. Dürren unterscheiden sich von den meisten anderen Gefahren dadurch , dass sie sich langsam , manchmal sogar über Jahre hinweg , entwickeln und ihr Beginn in der Regel schwer zu erkennen ist. Dürren sind nicht nur ein physikalisches Phänomen , da ihre Auswirkungen durch menschliche Aktivitäten und den Wasserbedarf noch verstärkt werden können. Dürren werden daher oft sowohl konzeptionell als auch operativ definiert. Operative Definitionen von Dürre , d. h. der Grad der Niederschlagsverringerung , der eine Dürre ausmacht , variieren je nach Ort , Klima und Umweltbereich .

```

Figure 9: Keywords and values used in the prompts for each hazard (2/3).

```

flood:

    hazard: Hochwasser

    keywords: Überschwemmung, Flut, Hochwasser, Überflutung, Flusshochwasser,
              Regenfälle, Sturzflut, Gletscherseeausbruch, Gletscherseeausbrüch,
              Gletschersee–Ausbruchsflut, Jahrhunderthochwasser

    hazard_event: Hochwasserereignisse

    definition: Ein allgemeiner Begriff für das Überlaufen von Wasser aus einem
                Flussbett auf normalerweise trockenes Land in der Aue (Flussüberschwemmung
                ), überdurchschnittlich hohe Wasserstände entlang der Küste (Küstenü
                berschwemmung) und in Seen oder Stauseen sowie Wasseransammlungen an oder
                in der Nähe des Ortes, an dem der Regen gefallen ist (Sturzfluten).

storm:

    hazard: Sturm

    keywords: Sturm, Stürm, Unwetter, Orkan, Blizzard, Derecho, Hagel, Zyklon,
              Gewitter, Tornado, Mikroburst, Hurrican, Hurrikan, Taifun, Blizzard

    hazard_event: Sturmereignis

    definition: Ein Sturm ist jeder gestörte Zustand der natürlichen Umwelt oder
                der Atmosphäre eines astronomischen Körpers. Er kann durch erhebliche Stö
                rungen der normalen Bedingungen gekennzeichnet sein, wie z. B. starker
                Wind, Tornados, Hagel, Donner und Blitz (Gewitter), starke Niederschläge (
                Schneesturm, Regensturm), starker Eisregen (Eissturm), starke Winde (
                tropischer Wirbelsturm, Sturm), Wind, der bestimmte Substanzen durch die
                Atmosphäre transportiert, wie z. B. bei einem Staubsturm, sowie andere
                Formen von Unwettern.

```

Figure 10: Keywords and values used in the prompts for each hazard (3/3).

Du bist ein Experte für die Klassifikation von Nachrichtenartikel bezüglich der Existenz von Referenzen auf \$hazard und extreme \$hazard_event.

Definition von \$hazard: \$definition

Synonyme für \$hazard: \$keywords.

Die Nachrichtenartikel müssen mit einem dieser Labels klassifiziert werden:

- Label 1: Das Dokument behandelt \$hazard, extreme \$hazard_event oder damit verbundene Auswirkungen.
- Label 0: Das Dokument hat KEINE Verbindung zu \$hazard oder extremen \$hazard_event.

Es sind nur konkrete, spezifische Naturereignisse in der realen Welt relevant. Artikel, die sich lediglich mit der Möglichkeit eines Ereignisses befassen, metaphorische Verwendungen, allgemeine Diskussionen über die Art der Gefahr oder Ereignisse sind nicht relevant.

Analysiere den Inhalt des Dokuments Satz für Satz sehr sorgfältig und vergib das Label 1 auch wenn nur ein Satz im Dokument relevant ist.

Entscheide, welches Label für diesen Nachrichtenartikel das richtige ist, und beginne deine Antwort entsprechend mit 0 oder 1.

Klassifiziere den folgenden Nachrichtenartikel:

\$text

Figure 11: Prompt used for the LLM experiment.

	1	2	3	4	5	6	7	8	9	10
landslide	98 see	bergbau	erdrutsch	wasser	rutschung	tagebau	gefahr	siedlung	erklären	bereich
	99 tal	felssturz	fels	stein	gestein	meter	kubikmeter	berg	stürzen	groß
	80 tote	erdrutsch	zahl	leiche	bergen	opfer	begraben	vermissen	vermisst	verschütten
	22 mensch	erdrutsch	leben	überschwemmung	haus	schwer	behörde	heftig	sterben	region
	60 haus	bewohner	unglück	erdrutsch	gebäude	wohnung	wohnen	früh	bürgermeisterin	ursache
wildfire	1 feuer	brand	waldbrand	flamme	waldbrand	hektar	feuerwehr	region	groß	kontrolle
	26 buschfeuer	australisch	koala	bundesstaat	kontinent	buschbrand	australier	tier	villa	ostküste
flood	43 mensch	überschwemmung	haus	leben	region	dpa	behörde	stadt	land	angabe
	20 hochwasser	sachsen	eibe	dresden	polen	pegel	meter	donau	brandenburg	tschechien
storm	16 unwetter	wasser	hochwasser	überschwemmung	heftig	region	schwer	regen	betreffen	schaden
	13 gewitter	wetter	regen	blitz	absagen	regnen	mark	himmel	heftig	schlecht
	7 hurrikan	sturm	bundesstaat	treffen	kilometer	wirbelsturm	land	schaden	windgeschwindigkeit	stunde
drought	457 dürre	notstand	ernsteausfall	vieh	regenzeit	ausmaß	zentrum	versicherung	wasserreservoir	helmholtz
	246 mitte	niedrigwasser	tag	fisch	wasser	stoff	fischsterben	sand	kreis	jugendliche
	319 trockenheit	waldbrandgefahr	feuchtigkeit	kanton	brandgefahr	ernteausfall	nässe	stress	notstand	leiter
heat wave	22 grad	temperatur	hitze	celsius	tag	hitzewelle	sommer	liegen	wetter	mensch
cold wave	1 mensch	sterben	leben	kältewelle	erfrieren	behörde	abdachlose	tote	zahl	duizend
	194 schneefall	heftig	stark	schneechaos	teil	fallen	sperrn	schneemasse	verkehr	blockieren
	137 eisregen	verspätung	reisend	glatt	mittag	behindern	glätte	vereist	schiene	störung
	12 winter	kalt	mild	wehen	östlich	wintermonat	stark	atlantik	flachland	luft
	111 kälte	warm	wärme	eisig	frieren	decke	kleidung	thermometer	anziehen	klirrend
	119 wetter	regen	kälteeinbruch	regnen	wetterlage	kalender	bauernregel	wetterstation	eishellig	schlecht

Table 6: Top 10 terms of each topic considered as relevant in the TM-B experiments. The second column contains an arbitrary topic ID.

Why Is This Green? LLM-Based Explanations of Implicit Green Practices in Social Media

Anna Glazkova^{1,2}, Olga Zakharova^{1,2}, Daria Lebedeva³

¹ Carbon Measurement Test Area in Tyumen' Region (FEWZ-2024-0016), Tyumen, Russia

² University of Tyumen, Tyumen, Russia

³ National Research University Higher School of Economics, Moscow, Russia
a.v.glazkova@utmn.ru, o.v.zakharova@utmn.ru, dlebedeva@hse.ru

Abstract

Identifying green practices in social media is not merely a matter of lexical matching. Many green practices are expressed implicitly, rely on shared background knowledge, or are embedded in broader contextual narratives. In this paper, we investigate how large language models (LLMs) explain expert annotations of green waste management practices and how they rationalize classification errors made by a fine-tuned model (mBART) on a Russian social media corpus (GreenRu). We analyze explanations generated by two LLMs (T-lite and GigaChat) in two settings: (1) explaining gold expert-assigned labels and (2) interpreting erroneous model predictions. Our qualitative and micro-quantitative analysis shows that green practices are frequently inferred through contextual reasoning rather than explicit terminology. Error patterns of mBART reveal overgeneralization, associative misinterpretation (e.g., linking food sharing to waste recycling), and detection of practices where none are present. We further compare explanatory strategies of the two LLMs. T-lite tends to rely on lexical cues and surface markers that may create an impression of a practice, while GigaChat more often reconstructs broader contextual interpretations. Expert feedback highlights limitations of formal textual analysis, sensitivity to missing contextual knowledge, and difficulties in aligning model reasoning with expert conceptual boundaries. Our findings suggest that explanation-based analysis is a productive tool for diagnosing classification errors and refining annotation guidelines. More broadly, the study demonstrates that modeling implicit sustainability discourse requires contextual grounding and deeper semantic integration beyond keyword-based approaches.

Keywords: green practices, social media, implicit information extraction, large language models, explainability, error analysis, environmental NLP

1. Introduction

The growing societal attention to environmental sustainability has led to the rapid expansion of online communities centered around ecological lifestyles, responsible consumption, and waste management. Social media platforms serve not only as spaces for information exchange but also as arenas where green practices are negotiated, normalized, and disseminated. Detecting such practices automatically is essential for large-scale analysis of grassroots environmental engagement, public discourse monitoring, and policy-oriented research.

However, identifying green practices in user-generated content is not a trivial text classification problem. Many environmentally relevant actions are expressed implicitly, rely on shared background knowledge, or are embedded in broader narratives about everyday life. For instance, bringing items to a named local initiative may constitute a practice of sharing or reuse, even if no explicit sustainability terminology is used. This makes green practice detection fundamentally a task of contextual interpretation rather than simple lexical matching.

Previous work in environmental natural language processing (NLP) has addressed related tasks such as climate change discourse analysis, environ-

mental stance detection, and sustainability-related information extraction (Beckles and Heidke, 2025; Peura et al., 2025; Zhang et al., 2024; Longo and Longo, 2025; Barz et al., 2025; Adam and Kogler, 2025). However, most studies focus either on topic modeling, sentiment and stance analysis, or explicit mention detection, rather than fine-grained behavioral practice identification—especially when practices are only indirectly expressed.

From a methodological perspective, our study intersects with research on multi-label text classification and explainability. Multi-label classification has been widely applied in various NLP domains. For example, Kementchedjnieva and Chalkidis (2023) demonstrated the effectiveness of encoder–decoder architectures for multi-label classification in biomedical and legal domains. Similarly, multilingual sequence-to-sequence models such as mBART (Tang et al., 2021) have proven competitive for complex classification tasks when reformulated as text generation.

Beyond classification performance, recent years have seen growing interest in explanation-based analysis of model behavior. Explainable AI in NLP has evolved from feature attribution methods (e.g., attention visualization, gradient-based saliency) toward natural language explanations generated by

LLMs (Zhao et al., 2024). Instruction-tuned LLMs can provide post-hoc rationalizations of predictions, enabling qualitative diagnosis of systematic errors (Brown et al., 2020; Kamoi et al., 2024). Such approaches are increasingly used not only to justify model outputs but also to probe implicit reasoning patterns and uncover annotation inconsistencies.

This work also relates to research on implicit information extraction and pragmatic inference in NLP. Detecting implicitly expressed events, relations, or intentions remains a longstanding challenge. Prior studies in implicit sentiment analysis, implicit hate speech detection, and implied event extraction show that models often rely on surface lexical associations rather than deeper semantic integration (ElSherief et al., 2021; Zhu et al., 2024; Loukachevitch et al., 2025; Masud et al., 2025). The difficulty increases in domain-specific contexts, where shared community knowledge plays a crucial role (Galli et al., 2025; Sengupta et al., 2025).

Within sustainability studies, the conceptualization of green practices as socially embedded behaviors rather than isolated keywords has been discussed in social science research (Zakharova et al., 2022). The GreenRu dataset (Zakharova and Glazkova, 2024) operationalizes this perspective by annotating concrete waste management practices at the sentence level in Russian-language social media posts. Unlike broader thematic datasets, GreenRu focuses on specific behavioral categories (e.g., waste sorting, sharing, refusing purchases), enabling fine-grained multi-label modeling.

In this paper, we move beyond standard performance evaluation and investigate how LLMs explain both correct expert annotations and systematic classification errors produced by a fine-tuned mBART classifier. We compare two Russian-language instruction-tuned LLMs: T-lite and GigaChat. Our analysis addresses three research questions:

1. How do LLMs reconstruct the reasoning behind expert-assigned green practice labels, especially when practices are expressed implicitly?
2. How do LLMs rationalize systematic classification errors, and what types of textual cues do they identify as misleading?
3. To what extent do explanation patterns reveal mismatches between expert conceptual boundaries and model-internal associations?

Our findings demonstrate that explanation-based analysis is a productive diagnostic tool for sustainability-oriented NLP tasks. Green practice detection emerges as a problem of contextual grounding and shared knowledge modeling rather than mere lexical classification. Moreover, comparing LLM-generated rationales with expert reflections

reveals not only model weaknesses but also latent ambiguities in annotation guidelines.

More broadly, the study contributes to the emerging intersection of environmental NLP and explainable AI by showing that modeling implicit sustainability discourse requires deeper semantic integration and explicit attention to contextual knowledge structures.

The remainder of the paper is structured as follows. Section 2 describes the dataset and task formulation, Section 3 presents the experimental setup, Section 4 reports the results, and Section 5 discusses implications of LLM-based explanations.

2. Dataset and Task

2.1. GreenRu Dataset

This study employs the GreenRu dataset (Zakharova and Glazkova, 2024)¹ to identify references to green waste management practices in Russian-language social media content. The dataset comprises 1,326 Russian-language posts, averaging 880 characters in length, sourced from online environmentally conscious communities.

The annotation of GreenRu was conducted at the sentence level using a multi-label scheme to capture the specific green waste practices mentioned, with sentences averaging 110 characters. The annotation framework is based on nine categories of green waste practices (Zakharova et al., 2022):

1. *waste sorting*, referring to the separation of waste by material type;
2. *studying product labeling* to identify packaging as a waste category;
3. *waste recycling*, the process of converting waste into reusable materials;
4. *signing petitions* as a means of influencing policy;
5. *refusing purchases* to minimize consumption and environmental impact;
6. *exchanging* unwanted items or services for desired ones;
7. *sharing* goods with others, either free or for a fee;
8. *participating in initiatives promoting responsible consumption*, such as workshops, festivals, or educational events;
9. *repairing* items instead of discarding them.

¹<https://github.com/green-solutions-lab/GreenRu>

The mentions of green practices are unevenly distributed. The most common practice is waste sorting (1,835 mentions), while the rarest is repairing (13 mentions).

2.2. Green Practice Detection

In this study, we explore the task of detecting mentions of green waste practices in social media. By green waste practices, we refer to concrete actions, behavioral patterns, or organizational decisions aimed at reducing environmental impact, conserving resources, promoting circular economy principles, or supporting broader environmental sustainability goals.

We formulate this problem as a multi-label text classification task. Given a textual fragment (typically a sentence or a short passage), the model must determine whether it mentions green practices.

3. Experimental Setup

3.1. Classification Model

For multi-label text classification, we used mBART² (Tang et al., 2021), a multilingual sequence-to-sequence transformer architecture introduced by Facebook AI Research. In previous studies (Glazkova and Zakharova, 2025), mBART achieved the best overall performance on green practice detection, outperforming both traditional machine learning approaches (e.g., kNN trained on sentence embeddings) and transformer-based models including BERT-based classifiers. A similar effectiveness of encoder-decoder models for multi-label text classification was demonstrated in the work of Kementchedjieva and Chalkidis (2023) for legal and biomedical texts.

Following the paper of Glazkova and Zakharova (2025), we fine-tuned mBART for 20 epochs using a maximum sequence length of 256 tokens. The model input consisted of sentences from the GreenRu dataset, and the output was a comma-separated list of the practices mentioned in the corresponding sentence.

3.2. LLMs for Explanation

For generating explanations, we used two instruction-based LLMs:

- T-lite³ (Stoianov et al., 2025), a Russian-language model built upon the Qwen 3 architecture, 8B params.

²<https://huggingface.co/facebook/mbart-large-50>

³<https://huggingface.co/t-tech/T-lite-it-2.1>

- GigaChat⁴ (Mamedov et al., 2025), a Russian-language model built on a Mixture-of-Experts (MoE) architecture with 10B total and 1.8B active parameters, it incorporates Multi-head Latent Attention (MLA) and Multi-Token Prediction (MTP) to achieve high inference throughput.

3.3. Explanation Settings

To evaluate the explanations generated by LLMs, we constructed two subsets of entries from the test portion of the GreenRu dataset. The dataset itself exhibits a naturally imbalanced distribution of practice types, reflecting the real-world prevalence of different green practices rather than an artificially balanced annotation scheme. Each subset contained 100 entries selected through randomized sampling while ensuring that all practice types were represented. Full class balance, however, could not be achieved due to the multi-label nature of the dataset: rare practices frequently co-occur with more common ones, making strict balancing infeasible without distorting the original data distribution.

The first subset (*random*) consisted of randomly selected entries together with their gold annotations of green practice mentions. The second subset (*error-only*) included entries for which none of the labels predicted by the mBART model matched the expert annotations. For these entries, we collected both the model predictions and the corresponding gold annotations. Examples of the selected entries and typical mBART errors are presented in Table 1. The normalized frequencies of gold-labeled practices in both subsets are shown in Figure 1. The figure demonstrates that while the overall distributions remain dominated by more frequent practices, both subsets preserve coverage of the full practice inventory, allowing a qualitative comparison of explanation behavior without substantially altering the underlying data distribution.

The prompts for analyzing expert annotations and mBART errors are presented in Table 2. The first prompt was intended to understand how human experts interpret implicit or explicit mentions of green practices, revealing the linguistic cues and contextual reasoning behind the gold labels (RQ1). The second prompt was designed to trace mBART errors back to specific textual elements, helping us identify whether mBART was misled by ambiguous wording, false positives (i.e. cases where the model erroneously detects a practice), or superficial patterns in the data (RQ2).

For both models, we used a generation temperature of 0.5. This value was chosen as a compromise between determinism and variability in generation. In preliminary experiments, lower values resulted

⁴<https://huggingface.co/ai-sage/GigaChat3-10B-A1.8B-bf16>

Text	Annotated practices (gold)	Annotated practices (mBART)
Упаковка с такой маркировкой не может быть переработана в принципе и закончит свой жизненный цикл на мусорном полигоне или в печи мусоросжигательного завода. { <i>Packaging with this labeling cannot be recycled at all and will end its life cycle in a landfill or in the furnace of an incineration plant</i> }	Изучение маркировки товаров { <i>Studying product labeling</i> }	Переработка отходов { <i>Waste recycling</i> }
В Великобритании изобрели #устройство для идентификации отходов - Recycling Identifying Device (R.I.D.). { <i>In the UK, a waste identification device called the Recycling Identifying Device (R.I.D.) has been invented</i> }	Сортировка отходов { <i>Waste sorting</i> }	Изучение маркировки товаров { <i>Studying product labeling</i> }
Фудшеринг - проект по спасению еды от утилизации, направленный на сохранение ресурсов нашей планеты и во имя борьбы с перепотреблением. { <i>Food sharing is a project aimed at saving food from disposal, preserving our planet's resources, and combating overconsumption</i> }	Обмен { <i>Sharing</i> }	Переработка отходов, отказ от покупок { <i>Waste recycling, refusing purchases</i> }

Table 1: Examples of entries and errors.

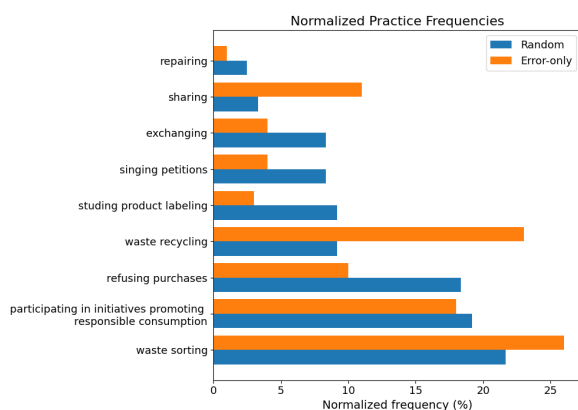


Figure 1: Normalized practice frequencies in the selected subsets.

in more repetitive explanations, while higher values increased variability without improving interpretability. The maximum number of tokens for explaining the expert annotations was set to 512, and for explaining the mBART annotations, it was set to 1024 tokens. In the second case, the maximum number of tokens was larger because the LLMs tended to generate longer responses to the second prompt. The actual energy consumption and carbon emissions of LLMs for 100 entries are given in Table 3.

4. Results

The average explanation length for the random subset was 1,346.67 characters for T-lite and 442.96 characters for GigaChat. The explanation length for gold labels in the error-only subset was 1,275.84 (T-lite) and 375.07 (GigaChat). The explanation length for mBART errors was 2,979.57 (T-lite) and 1,861.24 (GigaChat). On average, T-lite produces

longer and more lexically focused explanations, while GigaChat generates shorter, more abstractive justifications.

In the majority of cases, both LLMs explicitly refer to lexical units or short phrases from the input text as justification. However, T-lite more frequently relies on isolated lexical triggers, whereas GigaChat more often reconstructs larger semantic fragments (see Table 4). Both models often explicitly indicate whether a practice is expressed directly or implicitly. GigaChat more consistently signals implicitness (e.g., “the practice is not directly stated but inferred from...”), while T-lite more often assumes explicitness even when expert reasoning is contextual. In analyzing the errors of mBART, both models are generally able to identify plausible textual cues that could have triggered the incorrect prediction (RQ2). This suggests that many classification errors stem from systematic lexical or semantic associations rather than random noise.

The explanations produced by both LLMs were additionally analyzed in order to assess the reasoning strategies underlying their decisions. For each explanation column, we manually evaluated two indicators: (i) whether the model supported its decision with explicit lexical evidence by referring to fragments or linguistic cues present in the source text, and (ii) whether the model explicitly indicated that the described practice was expressed implicitly rather than directly stated. The analysis demonstrated that both models predominantly relied on textual grounding: lexical evidence was used in more than 95% of explanations across models. However, substantial differences were observed in how the models handled implicit meaning. The T-lite model identified practices as implicitly expressed in approximately 84% of cases,

Prompt ID	Prompt text
1	<p>The following sentence was annotated by experts as describing the following green practices: [PRACTICE_NAME] Sentence: "[TEXT]"</p> <p>Explain how exactly these practices are expressed in the text. If they are expressed implicitly, indicate what this interpretation is based on. If there is no explicit expression of these practices in the text, state this.</p>
2	<p>The model interpreted the following sentence as describing the following green practices: [PRACTICE_NAME] Sentence: "[TEXT]"</p> <p>Explain which elements of the text could have created the impression that these practices are being referred to. If this impression is weak or misleading, explain why.</p>

Table 2: Prompts used for generation. The prompts are given in English (translated from Russian). In cases where an entry had a single green practice annotated by experts or mBART, the corresponding singular phrasing was used.

Model	Prompt 1			Prompt 2		
	Time	Energy (kWh)	CO ₂ eq (g)	Time	Energy (kWh)	CO ₂ eq (g)
T-lite (8B)	0:28:45	0.1036	51.69	1:08:24	0.2425	120.93
GigaChat (1.8B active)	0:08:57	0.0287	7.28	0:40:35	0.1246	31.56

Table 3: The actual energy consumption and carbon emissions of LLMs for 100 entries using Google Colab Pro (NVIDIA A100-SXM4-40GB). The values are obtained with Carbontracker (Anthony et al., 2020).

whereas GigaChat did so considerably less frequently, in about 27% of explanations. In the analysis of model error interpretations, plausible textual triggers motivating incorrect predictions were identified in the clear majority of cases (over 90%), with only a small number of explanations lacking such triggers. These findings indicate that while both models ground their reasoning in the source text, they differ markedly in their tendency to infer implicit practices.

A central finding of our experiment is that green practice detection is fundamentally a task of contextual interpretation rather than lexical matching (RQ1). Many practices in GreenRu are expressed implicitly. For example, shared use initiatives (e.g., donation-based stores or exchange platforms) are not always described through explicit sustainability vocabulary. Human experts rely on background knowledge: they understand the institutional purpose of specific initiatives and interpret actions (e.g., bringing items to a named store) as part of a broader green practice. LLMs demonstrate partial sensitivity to this phenomenon. When contextual knowledge is available or widely encoded (e.g., widely known eco-initiatives), explanations become more accurate and coherent. When background knowledge is missing, models revert to surface-level lexical interpretation. In several cases, models explicitly mention “lack of context” or “insufficient

information” as a limitation. This meta-commentary indicates that LLMs can recognize epistemic uncertainty, yet they still tend to produce a forced interpretation.

Analyzing explanations for incorrect mBART predictions reveals systematic error types (RQ2). First, the model predicts a broader practice category instead of a more specific one (e.g., interpreting “bringing recyclables” as “recycling” rather than “waste sorting”). Second, closely related practices are confused due to associative links (e.g., food sharing interpreted as waste recycling because of lexical proximity between “food” and “waste”). Third, some keywords (e.g., “free,” “eco,” “volunteers”) activate sustainability interpretations even when the expert label depends on a more specific conceptual framing. Finally, the experts distinguish stages within a sustainability chain (label reading → sorting → transport → recycling), whereas the model treats them as a single undifferentiated process.

5. Expert Assessment of LLM Explanations

We invited green waste practice experts who participated in the original annotation process to review LLM-generated explanations. The expert assessment was conducted in a two-stage procedure.

Text	Вся макулатура будет переработана для вторичного использования. <i>{All waste paper will be recycled for reuse}</i>
Gold labels	Переработка отходов <i>{Waste recycling}</i>
mBART labels	Сортировка отходов <i>{Waste sorting}</i>
Explanations for the gold labels (Prompt 1)	
T-lite	<p>The practice of waste recycling is explicitly expressed in the sentence: "All waste paper will be recycled for reuse." Explanation: The text directly states: "will be recycled" — this is a verb denoting the action of recycling. It specifies what exactly is being recycled — waste paper (a type of waste based on paper and cardboard). It also indicates the purpose of the action — reuse, which aligns with the environmental goal of recycling: preventing waste from ending up in landfills, conserving resources, and reducing the ecological footprint. Conclusion: The practice is explicitly expressed through the verb "will be recycled" and the specification of the object (waste paper) and the purpose (reuse). Thus, the sentence fully corresponds to the description of the environmental practice of waste recycling.</p>
GigaChat	<p>In this sentence, the green practice of 'waste recycling' is explicitly expressed through the use of the verb 'will be recycled.' This indicates that all waste paper will undergo the recycling process for the purpose of reuse.</p>
Explanations for the mBART labels (Prompt 2)	
T-lite	<p>Excellent question. Let's examine why the sentence "All waste paper will be recycled for reuse" could be interpreted as describing an environmental practice, specifically waste sorting, and to what extent this impression is justified. Elements of the text that created the impression of "waste sorting": 1. The mention of "waste paper" — Waste paper consists of paper and cardboard waste that must be sorted during the collection and recycling stages. — The word "waste paper" itself indicates the separation of waste by type, i.e., sorting. 2. The phrase "will be recycled" — Waste recycling always begins with sorting. Without sorting materials by type (paper, plastic, glass, etc.), recycling is impossible or extremely inefficient. — This creates an association with a logical sequence: first sorting → then recycling. 3. The goal — "reuse" — This is a key goal of green practice aimed at resource conservation. — Sorting is one of the main ways to ensure the reuse of waste, especially waste paper. Conclusion: the impression of "waste sorting" is justified and strong Despite the fact that the sentence does not use the word "sorting", all the key elements point to the practice of waste separation: - waste paper → a type of waste that requires sorting - recycling → requires preliminary sorting - reuse → a goal achieved through sorting Therefore, the model's interpretation as "a description of an environmental practice — waste sorting" is justified and logical. Why does it not mislead? - There are no contradictions between the text and the interpretation. - All terms and logic correspond to real-world waste management practices. - Even if the emphasis is on recycling, it cannot be effective without sorting, especially for waste paper. Summary: Yes, the impression of "waste sorting" is not weak, but quite justified. The sentence clearly indicates an environmental practice related to sorting and does not mislead — on the contrary, it reflects the standard and correct approach to recycling waste paper. If the goal were, for example, "waste incineration" or "landfilling," the interpretation would be different. But here — everything points to an environmental practice with an emphasis on sorting and reuse.</p>
GigaChat	<p>The elements of the text that could create the impression that it refers to waste sorting: The word "waste paper" — This word is directly associated with the process of recycling paper and cardboard, which is part of the practice of waste sorting. The phrase "will be recycled" — Indicates the process of processing materials after their collection, which also relates to waste sorting. The expression "for reuse" — Emphasizes the idea of resource reutilization, which is a key aspect of environmental practice. Together, these elements create a clear understanding that the text refers to waste sorting and its subsequent recycling.</p>

Table 4: Examples of LLMs' output. The texts are given in English (translated from Russian).

First, one domain expert performed a detailed review of the model-generated explanations, providing structured comments on their adequacy, alignment with annotation guidelines, and typical reasoning patterns.

In the second stage, a second expert reviewed the initial assessment. In cases of uncertainty or disagreement, the experts discussed the instance jointly until a consensus was reached.

This procedure allowed us to ensure consistency of judgments while preserving detailed qualitative analysis. While the evaluation was not designed as a formal inter-annotator agreement study, the consensus-based review provides a reliable basis for identifying recurring strengths and limitations of

the explanations.

The goal of this section is to document how experts perceive model reasoning in relation to their own annotation principles.

5.1. Contextual Limitations of the Model

The main difficulty for LLMs is the lack of context, which directly affects its ability to reconstruct expert reasoning behind implicitly expressed practices (RQ1). If there is no information about the purpose of stores like "Mr. Thank You" ("Mr. Spasibo") and "Laugh-ka" ("Laffka"), it is very hard to understand why people bring their items there. While it is obvious to an expert that this is a practice of shared

use, the model starts to get confused, because it has to rely on the meaning of words in a short phrase, which does not always lead to correct conclusions. Another example is eco-friendly fashion. An expert knows how this practice addresses environmental problems and what established meanings are associated with it. On the other hand, the model can only guess what the word “eco-friendly” means in this phrase. When analyzing mistakes in such situations, LLMs point to the lack of context, insufficient information, and the formal nature of language analysis. The explanations of LLMs are especially accurate when it understands the context, for example, when it knows what is “plasticwatching” as an eco-practice.

5.2. Mismatch Between Expert Decisions and Model Reasoning

Another difficulty comes from expert decisions and highlights the mismatch between expert conceptual boundaries and model-internal associations (RQ3). For instance, experts decide that everything that happens before transporting separated waste to a recycler is waste sorting, which includes studying labels. The model, however, perceives the process as a whole: studying labels – sorting – transportation – recycling. It is difficult for it to distinguish where sorting ends and recycling begins, especially if there is no explicit statement about it. Another example of such an expert decision is the mention of volunteers. Experts assume that volunteers are involved in eco-practices, so their presence indicates the promotion of responsible consumption. At the same time, for the model volunteers are just people, so it needs additional information to make the conclusion. When analyzing expert judgment and model annotation, LLMs also do not have any information about the experts’ decisions, so they tend to agree with the model based on formal features.

For the analysis, LLMs refer to elements of the text in order to draw a conclusion about the validity of an expert’s or the model’s judgment. If it knows the context, it uses it.

5.3. Observed Weaknesses and Practical Value of Model Explanations

What can be concluded as drawbacks of the explanations of LLM:

1. The LLM approaches the analysis formally. For example, the word “free” is interpreted as refusing to buy something, because one cannot pay or purchase it;
2. LLMs may start analyzing elements of the text that are completely unrelated to the eco-

practice and are not significant from the expert’s point of view;

3. Complex actions, for example, the listing of the main activities of an environmental organization cannot be united by the LLM under a common meaning; it examines them separately;
4. LLMs do not always understand the task and begins to explain either the expert’s opinion or the model’s reasoning, pointing out what it considers weak, misleading, or, on the contrary, justified, or it proves the eco-friendliness of a practice rather than justifying its specific type;
5. LLMs will most likely make a mistake if it does not have access to all the information, for example, if it cannot see a photo of the packaging, it might conclude that the text is about studying labels, while there are no labels at the photo, only the specific type of the waste, such as polypropylene egg cartons;
6. LLMs may have overly high requirements for annotation, demanding exhaustive information about the practice, a full picture of it, which, of course, cannot be contained in a single phrase.

Why LLMs should be used for such tasks:

- From the point of view of the authors of the content, LLMs offer advice on how to convey a message more accurately, it helps to critically evaluate the wording in post texts, suggesting ways to make them more complete.
- From the expert’s point of view, LLMs help to identify the mistakes or inaccuracies in the annotation, see new interpretations of practices or to expand the range of practices, which are described in a given passage.
- From the programmer’s point of view, LLMs suggest how to fine-tune the classifier in order to avoid mistakes in the future, leading to the conclusion that training on the context is necessary.

6. Conclusion

The obtained results demonstrate that LLM explanations help reconstruct implicit expert reasoning (RQ1), diagnose systematic classification errors (RQ2), and reveal mismatches between expert annotation principles and model reasoning patterns (RQ3). Our findings are based on a diagnostic sample and should be interpreted as exploratory rather than statistically generalizable.

The analysis highlights that many errors stem from systematic semantic associations and overgeneralization rather than random noise. At the same time, LLM-generated explanations expose implicit assumptions embedded in both model predictions and expert annotations, thereby functioning as a diagnostic interface between computational models and domain knowledge.

More broadly, our study suggests that explanation-based analysis is a productive methodological tool for environmental NLP. By combining multi-label modeling with qualitative examination of generated rationales, researchers can better understand how sustainability discourse is represented, inferred, and sometimes misinterpreted by language models.

Future work may explore incorporating structured contextual knowledge into classification models, refining annotation guidelines based on explanation-driven feedback, and evaluating the faithfulness of generated explanations beyond their plausibility.

7. Acknowledgment

This study was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the Carbon Measurement Test Area in Tyumen' Region (FEWZ-2024-0016).

8. Bibliographical References

References

- Raven Adam and Marie Kogler. 2025. [Bidirectional topic matching: Quantifying thematic intersections between climate change and climate mitigation news corpora through topic modelling](#). In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 208–217, Vienna, Austria. Association for Computational Linguistics.
- Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. [Carbontracker: Tracking and predicting the carbon footprint of training deep learning models](#). ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. ArXiv:2007.03051.
- Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. [Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76, Tallinn, Estonia. University of Tartu Library.
- Valentina Tretti Beckles and Adrian Vergara Heidke. 2025. [Thematic categorization on pineapple production in Costa Rica: An exploratory analysis through topic modeling](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 44–55, Tallinn, Estonia. University of Tartu Library.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Carlo Galli, Anna V. Gavrilova, and Elena Calciolari. 2025. [Large language models in systematic review screening: Opportunities, challenges, and methodological considerations](#). *Information*, 16(5).
- Anna Glazkova and Olga Zakharova. 2025. [From data to grassroots initiatives: Leveraging transformer-based models for detecting green practices in social media](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 1–9, Tallinn, Estonia. University of Tartu Library.
- Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, et al. 2024. [Evaluating LLMs at detecting errors in LLM responses](#). *arXiv preprint arXiv:2404.03602*.
- Yova Kementchedjhieva and Ilias Chalkidis. 2023. [An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text](#). In *Findings of the Association for Com-*

- putational Linguistics: ACL 2023*, pages 5828–5843.
- Arianna Longo and Alessandro Y. Longo. 2025. [No AI on a dead planet: Sentiment and emotion analysis across Reddit communities on AI and the environment](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 77–83, Tallinn, Estonia. University of Tartu Library.
- Natalia Loukachevitch, Natalia Tkachenko, Anna Lapanitsyna, Mikhail Tikhomirov, and Nicolay Rusnachenko. 2025. [RuOpinionNE-2024: Extraction of opinion tuples from russian news texts](#). In *Proceedings of the International Conference "Dialogue"*, volume 2025.
- Valentin Mamedov, Evgenii Kosarev, Gregory Leytner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, Ainur Israfilova, Ivan Baskov, Artem Chervyakov, Emil Shakirov, Mikhail Kolesov, Daria Khomich, Daria Latortseva, Sergei Porkhun, Yury Fedorov, Oleg Kutuzov, Polina Kudriavtseva, Sofiia Soldatova, Kolodin Egor, Stanislav Pyatkin, Dzmitry Menshykh, Grafov Sergei Iurevich, Eldar Damirov, Vladimir Karlov, Ruslan Gaitukiev, Arkadiy Shatenov, Alena Fenogenova, Nikita Savushkin, and Fedor Minkin. 2025. [GigaChat family: Efficient Russian language modeling through mixture of experts architecture](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 93–106, Vienna, Austria. Association for Computational Linguistics.
- Sarah Masud, Ashutosh Bajpai, and Tanmoy Chakraborty. 2025. [Focal inferential infusion coupled with tractable density discrimination for implicit hate detection](#). *Natural Language Processing*, 31(6):1323–1349.
- Telma Peura, Attila Krizsán, Salla-Riikka Kuusalu, and Veronika Laippala. 2025. [Perspectives on forests and forestry in Finnish online discussions - a topic modeling approach to suomi24](#). In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 10–15, Tallinn, Estonia. University of Tartu Library.
- Saptarshi Sengupta, Wenpeng Yin, Preslav Nakov, Shreya Ghosh, and Suhang Wang. 2025. [Exploring language model generalization in low-resource extractive QA](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7106–7126, Abu Dhabi, UAE. Association for Computational Linguistics.
- Dmitrii Stoianov, Danil Taranets, Olga Tsymboi, Ramil Latypov, Almaz Dautov, Vladislav Kruglikov, Nikita Surkov, German Abramov, Pavel Gein, Dmitry Abulkhanov, et al. 2025. [T-pro 2.0: An efficient Russian hybrid-reasoning model and playground](#). *arXiv preprint arXiv:2512.10430*.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. [Multilingual translation from denoising pre-training](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- Olga Zakharova and Anna Glazkova. 2024. [GreenRu: A Russian dataset for detecting mentions of green practices in social media posts](#). *Applied Sciences*, 14(11):4466.
- Olga V Zakharova, Anna V Glazkova, Irina N Pupyshcheva, and Natalia V Kuznetsova. 2022. [The importance of green practices to reduce consumption](#). *Changing Societies & Personalities*. 2022. Vol. 6. Iss. 4, pages 884–905.
- Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024. [Granular analysis of social media users' truthfulness stances toward climate change factual claims](#). In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 233–240, Bangkok, Thailand. Association for Computational Linguistics.
- Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. [Explainability for large language models: A survey](#). *ACM Trans. Intell. Syst. Technol.*, 15(2).
- Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. [Can large language models understand context?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.

Introducing a Green Leaderboard for Sustainable Risk Prediction in Streaming NLP Shared Tasks.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Arturo Montejo-Ráez

Computer Science Department, SINAI, CEATIC, University of Jaén
Campus Las Lagunillas, 23071, Jaén, Spain
{amarmol, ammunoz, amontejo}@ujaen.es

Abstract

Current NLP shared-task evaluations predominantly rank systems by predictive performance, overlooking computational efficiency and environmental impact. This limitation is particularly critical in streaming and early risk detection scenarios, where models operate continuously, and resource consumption accumulates over time. We propose a sustainability-aware evaluation framework for streaming NLP tasks by introducing the Green Early Detection Score (GED), which integrates classification performance, detection timeliness, and carbon emissions. We also present an energy-based variant tailored to on-device early risk detection settings where energy consumption per inference is a key constraint. Applying these metrics to three editions (2023-2025) of the MentalRiskES shared task, we construct the first Green Leaderboard for early risk detection. Our results show that sustainability-aware ranking substantially reshapes system positions, highlighting efficient models that remain undervalued under performance-only evaluation.

Keywords: Natural Language Processing, Code Carbon, Energy Consumption, Environmental Impact, Large Language Models, Machine Learning

1. Introduction

The rapid advancement of Natural Language Processing (NLP) has led to increasingly powerful models, but this progress has come at a growing computational and environmental cost. Recent studies have shown that training large language models (LLMs) can emit amounts of CO₂ comparable to multiple transatlantic flights, raising serious concerns about the sustainability of current AI development practices (Strubell et al., 2019; Bender et al., 2021). As a result, the NLP community has begun to reflect on the environmental footprint of modern models. However, evaluation practices in shared tasks and competitions still overwhelmingly prioritize predictive performance, often relegating efficiency and sustainability indicators to secondary metadata.

Traditional leaderboards in NLP shared tasks rank systems exclusively according to a single predictive metric. In classification tasks, systems are ordered by Macro-F1, while regression tasks typically rely on RMSE. This single-metric paradigm creates implicit incentives to maximize performance regardless of computational efficiency, resource consumption, or deployment feasibility. These limitations are particularly pronounced in streaming scenarios such as early risk detection, where systems continuously process incoming data and must produce timely predictions. In such settings, computational and energy costs accumulate over time, yet current evaluation frameworks fail to account for these critical dimensions.

The MentalRiskES shared task (Mármol-Romero et al., 2023) constitutes a notable step forward in this regard. Focused on the early detection of mental health risks from Spanish social media, MentalRiskES has required participating teams to report detailed environmental indicators, including energy consumption and carbon emissions, alongside predictive metrics since 2023 (Mármol-Romero et al., 2024). This initiative has resulted in a unique multi-year dataset in which systems are evaluated not only in terms of accuracy and timeliness, but also with respect to their computational footprint. Nevertheless, despite the availability of this information, official rankings continue to be based solely on predictive performance, potentially favoring systems that achieve marginal gains at disproportionate computational or environmental cost.

Importantly, many early risk detection systems are explicitly designed for deployment on resource-constrained devices, such as smartphones or wearable platforms, where continuous monitoring, low latency, and limited battery capacity are key constraints. In these on-device scenarios, energy consumption per inference is often a more relevant and actionable metric than aggregate carbon emissions, as it directly impacts battery life, usability, and real-world feasibility. This observation motivates the need for evaluation metrics that explicitly account for energy efficiency and timeliness alongside predictive performance.

To address these limitations, we propose a sustainability-aware evaluation framework for streaming NLP tasks. We introduce the **Green**

Early Detection Score (GED), a composite metric that integrates predictive effectiveness, detection timeliness, and carbon emissions, and its energy-oriented variant, specifically tailored to on-device deployment scenarios. By applying these metrics to re-rank systems submitted to three editions of the MentalRiskES shared task, we construct the first *Green Leaderboard for early risk detection*. This leaderboard demonstrates how incorporating efficiency-aware criteria can significantly alter system rankings and encourage the development of models that are not only accurate and timely, but also practical and sustainable.

2. Related Work

The NLP and AI community has only recently begun systematically studying the energy and carbon costs of modern models. Early foundational work has shown that training large NLP models incurs substantial energy usage and financial costs, and recommends that researchers report resource metrics (e.g., training time) alongside accuracy (Strubell et al., 2019; Hershovich et al., 2022). This set the stage for “Green AI” efforts: Schwartz et al. (2020) surveyed multiple efficiency measures (CO2 emissions, electricity, FLOPs, etc.) and argued for explicitly treating performance-compute trade-offs as a metric, for example by plotting accuracy against training size or energy use. In this spirit, Bender et al. (2021) famously warned of the dangers of “ever-larger” language models. They urge weighing the environmental and financial costs first in model design, rather than blindly scaling up, recommending more careful data curation and exploring research beyond ever-larger models. In short, these works emphasize that, in NLP and AI, carbon costs should be reported and minimized as part of standard practice.

To quantify these effects, several tools and frameworks have been developed. Bannour et al. (2021) surveyed six popular carbon-tracking tools (CarbonTracker, Experiment-Impact-Tracker, Green Algorithms, ML-CO2 Impact, etc.) and applied them to NLP experiments. They demonstrate that tools differ in scope and methodology but all aim to measure kWh and CO2 of training runs. Lannelongue et al. (2021) went further by providing a simple online calculator that estimates carbon footprint from compute hardware, runtime, and location; they applied it to NLP tasks. Their calculator¹ and open methodology have been adopted widely, illustrating how to generalize carbon accounting across domain. In practice, researchers now often use tools like CodeCarbon (Schmidt et al., 2021) to record emissions during experiments. These measurement techniques are now being applied within

¹<https://www.green-algorithms.org/>

shared evaluation campaigns. Mármol-Romero et al. (2024) analyze the MentalRiskES shared task on Spanish mental health risk detection, where organizers required participants to submit carbon-efficiency metrics alongside predictions. This study empirically correlates submitted CO2 emissions with model architecture and performance, demonstrating how a shared-task framework can reveal the ecological footprint of diverse NLP systems. In short, a growing number of NLP competitions now include energy reporting, enabling analyses of model emissions at scale (Vallecillo-Rodríguez et al., 2024).

Beyond tools, researchers have also articulated conceptual frameworks for “sustainable AI”. Van Wynsberghe (2021) defines Sustainable AI as AI that is compatible with maintaining environmental resources for current and future generations. More concretely, Bolón-Canedo et al. (2024) review the notion of “Green AI”, distinguishing green-by-AI (using AI to aid environmental applications) from green-in-AI (making AI itself more efficient). They highlight strategies like designing energy-efficient models, using renewable-energy data centers, and adding regulatory incentives. Similarly, Kaack et al. (2022) argues for aligning AI with climate change mitigation, calling for high-level commitments and policies to ensure AI development reduces rather than worsens emissions. In practice, many researchers promote concrete best practices (e.g. using efficient hardware, pruning, distillation) and transparency: reporting climate-related performance is urged as a way to drive improvements (Hershovich et al., 2022; Henderson et al., 2020).

3. Methodology

3.1. Task Setting

MentalRiskES is formulated as an online early risk detection task. For each user u , systems receive a temporally ordered sequence of messages:

$$\{m_{u,1}, m_{u,2}, \dots, m_{u,T}\}$$

At each round t , the system must emit a prediction regarding whether the user is at risk of a given mental disorder. Predictions are produced incrementally, simulating a real-time monitoring scenario where decisions must be made under partial information.

Let t_u^* denote the first round at which a system correctly detects the disorder for user u . If the system never detects the disorder, t_u^* is undefined and penalized according to the official evaluation protocol. MentalRiskES evaluates systems along two main dimensions:

- **Predictive performance**, measured using Macro-F1.
- **Detection timeliness**, measured using Early Risk Detection Error (ERDE) (Losada and Crestani, 2016). (ERDE). ERDE combines correctness and delay: false negatives receive the maximum penalty, while correct positive detections become increasingly costly as the system consumes more messages before emitting the alert. In MentalRiskES, ERDE30 uses a delay parameter of 30 messages, so late but correct detections are penalized more strongly than early correct alerts.

3.2. Environmental Impact Measurement

In addition to predictive metrics, MentalRiskES requires participants to report environmental metadata for each prediction round, measured using CodeCarbon (Courty et al., 2024). For each system and prediction round, they collect:

- CO₂ emissions in kilograms of CO₂ equivalent.
- Energy consumption in kilowatt-hours.
- Hardware configuration, including CPU, GPU, and RAM usage.

3.3. Data Collection

We analyze all systems submitted to the MentalRiskES shared tasks during the 2023, 2024, and 2025 editions. For each system, we extract:

- Predictive metrics: Macro-F1 and ERDE30 (metrics used for the official ranking).
- Environmental metrics: average and cumulative CO₂ emissions and total energy consumption.
- System characteristics: model family (machine learning, deep learning, or LLMs) and reported hardware configuration.

This results in a unified dataset covering more than 170 systems evaluated under identical streaming conditions. Data from the 2023 and 2024 editions are based on the MentalRiskES dataset (Mármol Romero et al., 2024), while data from the 2025 edition correspond to the PRECOM dataset (Álvarez-Ojeda et al., 2025).

4. Green Early Detection Evaluation Framework

To integrate predictive effectiveness, detection timeliness, and computational efficiency into a single

evaluation framework, we adopt a normalized multi-criteria approach rather than a single ad-hoc scalar metric. This design choice follows established practices in Green AI and multi-objective evaluation, where heterogeneous dimensions are first normalized and then combined transparently.

Predictive Effectiveness Score We define a unified predictive effectiveness score that jointly captures classification performance and detection timeliness:

$$P = \text{Macro-F1} \cdot (1 - \text{ERDE}_{30}) \quad (1)$$

This formulation ensures that systems are rewarded only when they achieve both high predictive accuracy and early detection. Systems that detect risks late or inconsistently are penalized, even if their final classification performance is strong. This reflects the clinical motivation of early intervention scenarios, where timeliness is a first-class requirement.

The resulting score is bounded in the interval $[0, 1]$, facilitating comparison and aggregation with other normalized indicators.

Efficiency Score To account for computational efficiency, we normalize environmental indicators across all evaluated systems. Let X denote an efficiency-related metric, such as mean CO₂ emissions per prediction or mean energy consumption per prediction. We compute a normalized efficiency score as:

$$E_X = 1 - \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

where X_{\min} and X_{\max} correspond to the minimum and maximum observed values across all submissions. This transformation yields an efficiency score in $[0, 1]$, where higher values indicate more efficient systems.

Importantly, efficiency scores are *not comparable across different years*, as each edition of the shared task relies on distinct datasets, experimental conditions, and submission pools. Consequently, normalization is performed independently for each year to ensure fair and meaningful comparisons only among systems evaluated under the same conditions.

This normalization strategy offers three advantages. First, it avoids mixing heterogeneous physical units directly with predictive metrics. Second, it constrains all scores to a common bounded range $[0, 1]$, making them directly composable with the predictive effectiveness score P without requiring further scaling; scores derived from raw physical units (e.g., kWh or kg CO₂eq) would otherwise dominate the composite metric due to differences

in magnitude. Third, it allows efficiency to be evaluated relative to the actual design space explored by participants in a given year, which is particularly appropriate in shared-task settings. We acknowledge that this relative normalization means scores depend on the composition of a given year’s submission pool; a system’s efficiency score may shift if the pool changes. This is a deliberate design choice: the goal is to rank systems fairly within a shared experimental context, not to assign absolute efficiency values independent of competition conditions.

Green Early Detection Score We combine predictive effectiveness and efficiency into a GED score using a weighted linear aggregation:

$$\text{GED} = \alpha P + \beta E_X, \quad \text{with } \alpha + \beta = 1 \quad (3)$$

The weights α and β explicitly control the trade-off between predictive effectiveness and efficiency. In our analysis, we consider two configurations:

- Performance-oriented Green Score: $\alpha = 0.7$, $\beta = 0.3$
- Balanced Green Score: $\alpha = 0.5$, $\beta = 0.5$

By making the weighting scheme explicit, this framework avoids hidden assumptions and allows organizers and practitioners to adapt the evaluation to different deployment priorities. Importantly, the resulting score remains bounded, interpretable, and suitable for leaderboard-based ranking. We emphasize that the GED is not intended to replace traditional performance metrics, but to complement them by making efficiency an explicit and actionable evaluation dimension.

4.1. Energy and Carbon-Aware Instantiations

Sustainability constraints vary substantially depending on the intended deployment context. In centralized evaluation settings, aggregate carbon emissions are often the primary concern, while in continuous and on-device monitoring scenarios, energy consumption per inference directly impacts battery life, latency, and usability.

To reflect these differences, we instantiate the proposed framework using two complementary efficiency indicators.

Carbon-Aware Green Score For evaluating global environmental impact, we define a carbon-aware Green Early Detection Score by setting $X = C$, where C denotes the mean CO₂-equivalent emissions per prediction:

$$E_{\text{CO}_2} = 1 - \frac{C - C_{\min}}{C_{\max} - C_{\min}} \quad (4)$$

$$\text{GED}_{\text{CO}_2} = \alpha P + \beta E_{\text{CO}_2} \quad (5)$$

This score favors systems that achieve strong early detection performance while minimizing carbon emissions, encouraging environmentally responsible model design at the system level.

Energy-Aware Green Score For on-device and real-time deployment scenarios, we instead consider mean energy consumption per prediction E as the relevant efficiency indicator:

$$E_{\text{energy}} = 1 - \frac{E - E_{\min}}{E_{\max} - E_{\min}} \quad (6)$$

$$\text{GED}_{\text{energy}} = \alpha P + \beta E_{\text{energy}} \quad (7)$$

This formulation directly rewards models that provide high-quality early detection under strict energy constraints. Notably, systems optimized for low carbon emissions are not necessarily optimal in terms of per-inference energy usage, making this distinction critical for realistic deployment assessment.

By maintaining a unified evaluation framework and varying only the efficiency indicator, we ensure methodological consistency while enabling context-aware sustainability evaluation. Rather than proposing a single universal metric, our approach supports multiple, transparent instantiations aligned with different operational priorities.

5. Green Leaderboard

The Green Leaderboard re-ranks all participating systems from the 2023, 2024, and 2025 editions of the MentalRiskES shared task using the proposed GED metric, incorporating both predictive performance and environmental efficiency. The following observations can be drawn from Tables 1, 2, and 3.

Comparing the official rankings, which are based solely on Macro-F1, with the GED-based rankings reveals notable differences. Several systems that did not occupy top positions in the traditional leaderboard achieve leading positions under GED, highlighting the importance of considering sustainability and efficiency alongside predictive effectiveness. This trend is particularly evident for lightweight machine learning (ML) models and compact neural architectures, which maintain competitive early detection performance while consuming significantly less energy and producing lower CO₂ emissions.

For most teams, GED_CO2 and GED_energy values are closely aligned, indicating that systems optimized for carbon efficiency are generally also energy-efficient per prediction. However, in more

complex tasks (e.g., Task 2c in 2023), discrepancies between GED_CO2 and GED_energy are observed. Some systems achieve high carbon efficiency but relatively lower energy efficiency, demonstrating that these two indicators capture complementary aspects of environmental performance.

Yearly Trends

- **2023:** CO₂ and energy rankings are largely aligned in Task 1a, whereas differences become more pronounced in Tasks 2c and 3a, highlighting the amplification of efficiency effects in complex tasks.
- **2024:** Teams such as ELiRF-UPV and UnibucAI consistently appear at the top of the Green Leaderboard, demonstrating that medium-sized models can compete with larger architectures when sustainability is considered.
- **2025:** New teams such as MCDI and PUXai appear among the top-ranked systems. Variability between GED_CO2 and GED_energy increases, indicating that optimizing solely for one environmental indicator does not necessarily maximize the overall GED score.

6. Sustainability Analysis

This section analyzes the relationship between predictive effectiveness, model complexity, and resource consumption across all systems submitted to the MentalRiskES shared tasks from 2023 to 2025. Rather than focusing on individual rankings, we examine global trends that emerge when sustainability indicators are considered jointly with performance metrics.

Performance vs. Energy Consumption Figure 1 illustrates the relationship between Macro-F1 and mean energy consumption per prediction across all submitted systems. A clear trade-off emerges: while some high-performing systems achieve strong Macro-F1 scores, they do so at substantially higher energy costs, often spanning several orders of magnitude.

Notably, the Pareto frontier reveals that competitive performance can be obtained with relatively low energy consumption. Several lightweight ML models and hybrid approaches (e.g., Transformer + ML) lie on or near the frontier, achieving Macro-F1 values comparable to larger transformer-based systems while consuming significantly less energy per inference. This suggests diminishing returns in performance when increasing model complexity beyond a certain point, particularly in continuous streaming settings.

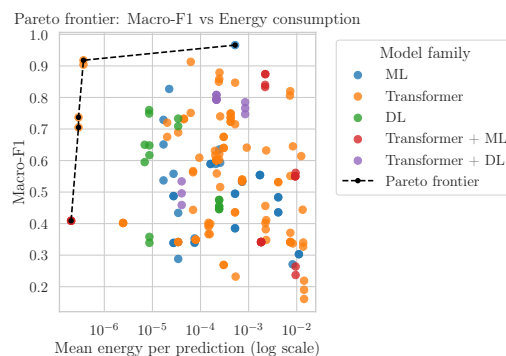


Figure 1: Pareto frontier between predictive performance (Macro-F1) and mean energy consumption per prediction (log scale) across all MentalRiskES submissions from 2023 to 2025. Points are grouped by model family. The frontier highlights systems that achieve competitive performance under strict energy constraints.

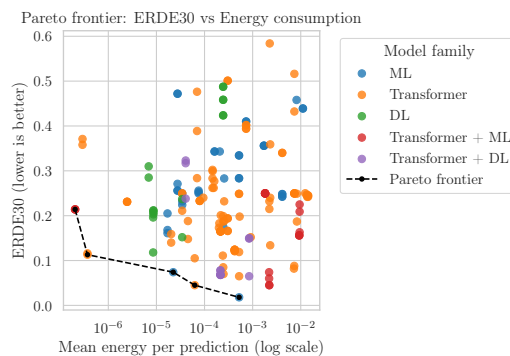


Figure 2: Pareto frontier between detection timeliness (ERDE30, lower is better) and mean energy consumption per prediction (log scale). The figure shows that early risk detection can be achieved without resorting to energy-intensive models.

Figure 2 further examines energy consumption in relation to ERDE30. Systems with lower energy usage tend to exhibit a wider range of timeliness behavior, but the Pareto frontier again highlights that early detection performance does not require energy-intensive architectures. Importantly, several energy-efficient systems achieve low ERDE30 values, indicating that prompt detection is compatible with strict energy constraints.

Performance vs. Emissions Figures 3 and 4 present analogous trends when mean CO₂ emissions per prediction are considered instead of energy consumption. As with energy, emissions span multiple orders of magnitude, reflecting substantial variability in hardware choices and computational strategies across teams.

Table 1: Top 10 Green Leaderboard for all tasks (2023) showing both GED_CO2 and GED_energy. The “Official Rank” column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while “CO2 Rank” and “Energy Rank” refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	P ↑	GED_CO2 ↑	GED_energy ↑
Task 1a									
1	1	1	CIMAT-NLP-GTO	0	0.966	0.018	0.949	0.952	0.953
2	2	3	UNSL	1	0.913	0.045	0.872	0.909	0.909
4	3	2	UMUTeam	0	0.918	0.113	0.814	0.870	0.870
3	4	5	VICOM-nlp	2	0.879	0.070	0.817	0.870	0.867
5	5	4	UMUTeam	1	0.904	0.116	0.799	0.859	0.859
6	6	6	VICOM-nlp	1	0.859	0.085	0.786	0.848	0.845
7	7	8	CIMAT-NLP-GTO	1	0.847	0.065	0.792	0.843	0.843
8	8	9	plncmm	0	0.827	0.074	0.766	0.836	0.836
9	9	7	VICOM-nlp	0	0.850	0.111	0.756	0.827	0.824
11	11	12	NLP-UNED	0	0.760	0.118	0.670	0.769	0.769
Task 2a									
1	1	2	UNSL	1	0.733	0.148	0.625	0.736	0.736
2	2	5	SINAI-SELA	0	0.720	0.140	0.619	0.733	0.733
3	3	4	TextualTherapists	1	0.729	0.161	0.612	0.728	0.728
4	4	3	UNSL	0	0.731	0.188	0.594	0.715	0.715
5	5	7	SINAI-SELA	1	0.675	0.159	0.568	0.697	0.697
6	6	8	TextualTherapists	0	0.651	0.168	0.542	0.679	0.679
7	7	12	VICOM-nlp	2	0.631	0.173	0.522	0.664	0.661
8	8	11	CIMAT-NLP-GTO	0	0.635	0.175	0.524	0.661	0.662
9	9	9	NLP-UNED	1	0.648	0.207	0.514	0.660	0.660
10	11	15	VICOM-nlp	1	0.616	0.183	0.503	0.651	0.648
Task 2c									
1	1	1	NLP-UNED	1	0.358	0.203	0.285	0.500	0.500
2	2	2	NLP-UNED	0	0.339	0.211	0.267	0.487	0.487
3	3	3	plncmm	0	0.288	0.232	0.221	0.454	0.454
4	4	4	I2C-UHU	0	0.232	0.198	0.186	0.422	0.419
5	5	5	SPIN	1	0.219	0.242	0.166	0.291	0.117
6	6	6	SPIN	0	0.190	0.245	0.143	0.276	0.101
7	7	7	SPIN	2	0.161	0.245	0.122	0.260	0.085
Task 3a									
1	1	1	CIMAT-NLP-GTO	2	0.740	0.188	0.601	0.715	0.715
2	2	2	NLP-UNED	1	0.650	0.285	0.465	0.625	0.625
3	3	5	CIMAT-NLP-GTO	0	0.593	0.283	0.425	0.592	0.592
4	4	4	NLP-UNED	0	0.595	0.310	0.411	0.587	0.587
5	5	6	CIMAT-NLP-GTO	1	0.516	0.232	0.396	0.572	0.572
6	6	8	UPM	0	0.402	0.231	0.309	0.516	0.516
6	6	8	UPM	2	0.402	0.231	0.309	0.516	0.516
6	6	8	UPM	1	0.402	0.231	0.309	0.516	0.516
9	9	3	CIMAT-NLP	0	0.614	0.250	0.460	0.342	0.361
10	10	7	CIMAT-NLP	1	0.444	0.247	0.334	0.234	0.254

The Macro-F1 vs. emissions plot (Figure 3) shows that high predictive performance is not exclusive to high-emission systems. Several low-emission models achieve performance levels close to the best-performing systems, again forming a well-defined Pareto frontier. This indicates that environmentally efficient solutions remain competitive in terms of accuracy.

Similarly, Figure 4 demonstrates that early detection timeliness does not systematically improve with higher emissions. In fact, many of the lowest-ERDE systems operate at relatively low emission levels, reinforcing the conclusion that carbon efficiency and early detection objectives are not inherently conflicting.

Temporal Evolution of Model Families Figure 5 shows the evolution of model family adoption across the 2023–2025 editions of the shared task. While transformer-based architectures remain prominent throughout all years, their dominance decreases slightly over time, coinciding with an increased presence of hybrid and lightweight machine learning approaches.

This trend suggests a gradual shift toward more efficiency-aware system design. In particular, the growing proportion of Transformer + ML systems in-

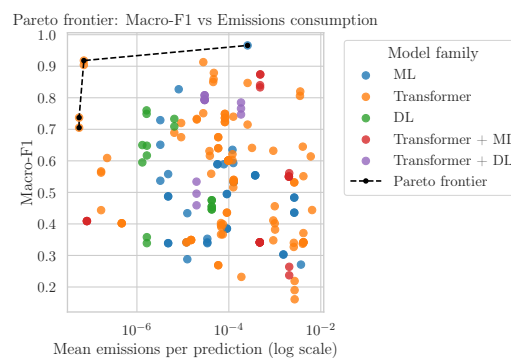


Figure 3: Pareto frontier between predictive performance (Macro-F1) and mean CO₂ emissions per prediction (log scale). Several low-emission systems achieve performance comparable to high-emission models, illustrating the trade-off between accuracy and environmental impact.

icates that teams increasingly combine representational power with computational efficiency, likely motivated by both sustainability concerns and deployment constraints.

Table 2: Top 10 Green Leaderboard for all tasks (2024) showing both GED_CO2 and GED_energy. The “Official Rank” column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while “CO2 Rank” and “Energy Rank” refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	F ↑	GED_CO2 ↑	GED_energy ↑
Task 1									
1	1	1	ELIRF-UPV	2	0.874	0.045	0.835	0.845	0.824
2	2	4	UnibucAI	0	0.808	0.078	0.745	0.819	0.816
3	3	5	UnibucAI	1	0.795	0.069	0.740	0.816	0.812
4	4	6	UnibucAI	2	0.793	0.068	0.739	0.815	0.812
5	6	3	ELIRF-UPV	0	0.833	0.060	0.783	0.809	0.789
6	7	2	ELIRF-UPV	1	0.840	0.074	0.778	0.806	0.785
7	5	7	UNED-GELP	0	0.785	0.065	0.734	0.799	0.791
8	8	9	lxa-Med	1	0.749	0.124	0.656	0.753	0.748
9	9	11	lxa-Med	2	0.736	0.121	0.647	0.746	0.741
10	10	8	UNED-GELP	2	0.766	0.149	0.652	0.741	0.733
Task 2									
1	1	1	ELIRF-UPV	0	0.874	0.045	0.835	0.845	0.824
2	2	2	UnibucAI	2	0.808	0.078	0.745	0.819	0.816
4	4	3	UnibucAI	1	0.793	0.068	0.739	0.815	0.812
3	3	3	UnibucAI	0	0.793	0.068	0.739	0.815	0.812
5	5	5	lxa-Med	1	0.749	0.124	0.656	0.753	0.748
6	6	6	lxa-Med	2	0.736	0.121	0.647	0.746	0.741
7	7	7	lxa-Med	0	0.723	0.124	0.633	0.737	0.732
8	8	8	UMUTeam	2	0.675	0.166	0.563	0.689	0.686
9	9	9	UMUTeam	0	0.640	0.194	0.516	0.656	0.653
10	10	10	UC3M-DAD	0	0.601	0.165	0.502	0.643	0.645
Task 3									
1	1	1	UnibucAI	0	0.534	0.238	0.407	0.583	0.584
2	2	2	UnibucAI	1	0.496	0.317	0.339	0.536	0.536
3	3	5	V team	0	0.409	0.214	0.321	0.525	0.525
3	3	5	V team	2	0.409	0.214	0.321	0.525	0.525
3	3	5	V team	1	0.409	0.214	0.321	0.525	0.525
6	6	3	UnibucAI	2	0.459	0.323	0.311	0.516	0.516
7	7	4	UNED-GELP	0	0.456	0.215	0.358	0.470	0.490
8	8	8	UNED-GELP	1	0.402	0.232	0.309	0.435	0.456
9	9	9	UNED-GELP	2	0.382	0.584	0.159	0.329	0.350

Table 3: Top 10 Green Leaderboard for all tasks (2025) showing both GED_CO2 and GED_energy. The “Official Rank” column refers to the official performance-based rank from the original MentalRiskES leaderboard (based on Macro-F1), while “CO2 Rank” and “Energy Rank” refer to positions under the GED_CO2 and GED_energy metrics, respectively. Best values per column are highlighted in bold.

CO2 Rank	Energy Rank	Official Rank	Team	Run	Macro_F1 ↑	ERDE30 ↓	F ↑	GED_CO2 ↑	GED_energy ↑
Task 1									
1	1	2	UNSL	0	0.563	0.284	0.403	0.582	0.580
2	2	1	UNSL	2	0.567	0.389	0.346	0.542	0.541
3	3	8	UC3Mental	1	0.495	0.334	0.329	0.524	0.510
4	4	15	UC3Mental	2	0.436	0.249	0.327	0.523	0.509
5	5	5	ELIRF-UPV	2	0.534	0.394	0.324	0.517	0.498
6	6	4	ELIRF-UPV	1	0.540	0.402	0.323	0.517	0.497
7	9	6	ELIRF-UPV	0	0.533	0.410	0.314	0.511	0.491
8	7	18	PUXai	2	0.396	0.283	0.284	0.494	0.494
9	8	17	PUXai	0	0.403	0.302	0.281	0.492	0.492
10	19	19	UC3Mental	0	0.385	0.283	0.276	0.487	0.473
Task 2									
1	1	4	HULAT_UC3M	1	0.558	0.271	0.407	0.584	0.585
3	3	1	MCDI	0	0.589	0.343	0.387	0.567	0.565
2	2	1	MCDI	1	0.589	0.343	0.387	0.567	0.565
4	4	1	MCDI	2	0.589	0.343	0.387	0.565	0.564
5	5	9	UC3Mental	1	0.495	0.334	0.329	0.524	0.510
6	6	15	UC3Mental	2	0.436	0.249	0.327	0.523	0.509
8	8	5	ELIRF-UPV	1	0.540	0.402	0.323	0.517	0.497
7	7	6	ELIRF-UPV	2	0.534	0.394	0.324	0.517	0.498
9	10	7	ELIRF-UPV	0	0.533	0.410	0.314	0.511	0.491
10	9	17	PUXai	2	0.399	0.277	0.289	0.497	0.497

Biases Induced by Ranking Figure 6 illustrates the model family composition of the Top-10 systems ranked exclusively by Macro-F1. Across all years, this ranking strategy disproportionately favors transformer-based and deep learning architectures, despite their often higher energy and emission costs.

In contrast, Figures 7 and 8 show the Top-10 composition under the carbon-based and energy-based GED rankings, respectively. When sustainability criteria are incorporated, the leaderboard becomes markedly more diverse, with a substantial increase in ML and hybrid models.

The effect is particularly pronounced under the energy-based GED, where lightweight models consistently replace more computationally intensive systems in the Top-10. This highlights how performance-only evaluation implicitly biases shared-task outcomes toward resource-heavy solutions, whereas sustainability-aware metrics expose a broader and more realistic design space.

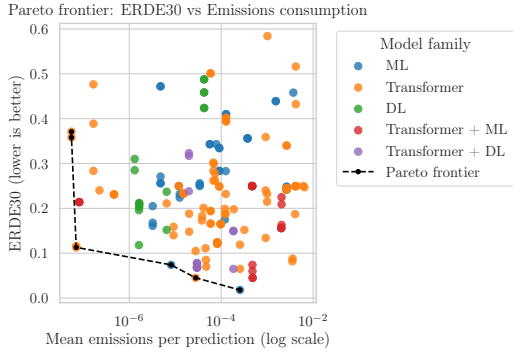


Figure 4: Pareto frontier between detection timeliness (ERDE30, lower is better) and mean CO₂ emissions per prediction (log scale). Results indicate that improved timeliness does not necessarily require higher carbon emissions.

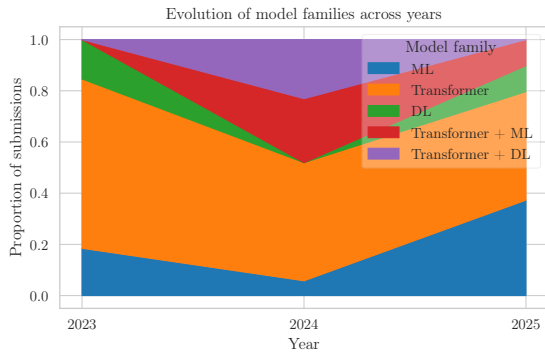


Figure 5: Temporal evolution of model family distribution across the 2023, 2024, and 2025 editions of the MentalRiskES shared task. While transformer-based models remain prevalent, hybrid and lightweight approaches gain prominence over time.

6.1. Implications for Streaming NLP Evaluation

Overall, the sustainability analysis demonstrates that evaluation metrics directly shape the types of models that are incentivized and rewarded. Performance-only leaderboards systematically favor computationally intensive systems, even when their advantages over more efficient alternatives are marginal. By contrast, sustainability-aware metrics reveal a broader and more realistic design space, in which multiple architectures achieve competitive early detection performance under strict resource constraints.

These findings support the inclusion of explicit efficiency-aware criteria in streaming NLP evaluations. Making sustainability visible at the ranking level encourages responsible model development and helps align shared-task outcomes with real-

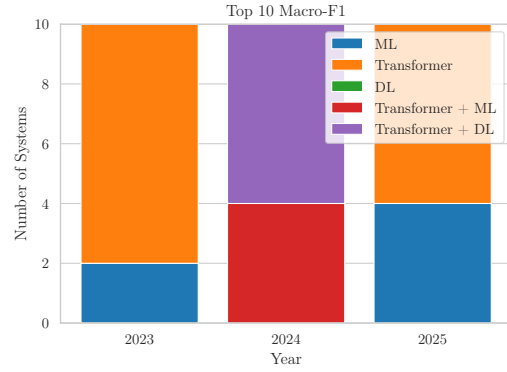


Figure 6: Model family composition of the Top-10 systems ranked exclusively by Macro-F1 for each year. Performance-only ranking disproportionately favors computationally intensive architectures.

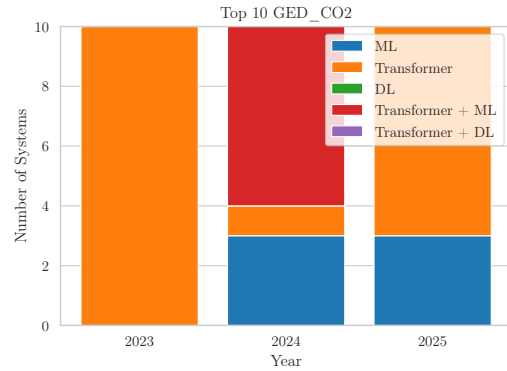


Figure 7: Model family composition of the Top-10 systems ranked using the carbon-based Green Early Detection Score (GED_{CO2}). Incorporating emissions into the ranking increases the representation of lightweight and hybrid models.

world deployment requirements, particularly in sensitive domains such as mental health monitoring.

7. Conclusion

In this paper, we introduced a sustainability-aware evaluation framework for streaming NLP tasks, addressing the limitations of performance-only leaderboards in early risk detection scenarios. We proposed the GED Score, a flexible and transparent metric that jointly accounts for predictive performance, detection timeliness, and environmental efficiency, with both carbon- and energy-aware instantiations.

By applying this framework to three consecutive editions (2023–2025) of the MentalRiskES shared task, we constructed the first Green Leaderboard for early mental health risk detection. Our analysis shows that incorporating sustainability criteria substantially reshapes system rankings, frequently

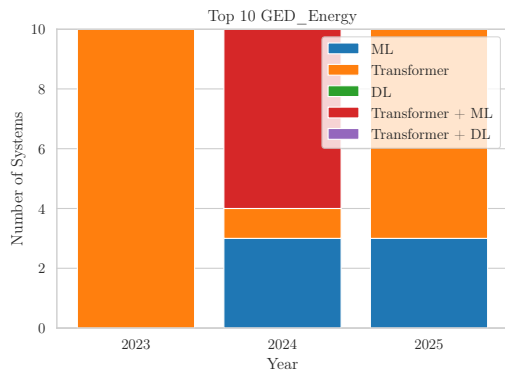


Figure 8: Model family composition of the Top-10 systems ranked using the energy-based Green Early Detection Score (GED_{Energy}). Energy-aware evaluation further amplifies the presence of efficient model families.

elevating lightweight and hybrid approaches that remain competitive in predictive effectiveness while dramatically reducing energy consumption and CO_2 emissions.

Importantly, we show that evaluation metrics directly influence the types of models incentivized in shared-task settings. Performance-only evaluation implicitly favors resource-intensive architectures, whereas sustainability-aware metrics expose a broader and more realistic design space aligned with real-world deployment constraints. This is particularly critical in sensitive domains such as mental health, where systems are expected to operate continuously, often on resource-constrained devices.

We argue that sustainability-aware evaluation is both feasible within existing shared-task infrastructures and necessary to encourage responsible model development. As future work, we plan to extend the proposed framework by incorporating additional efficiency indicators, such as inference latency and memory footprint, and by exploring normalization strategies that account for heterogeneous hardware configurations. We believe that making efficiency an explicit and first-class evaluation dimension is essential for advancing environmentally responsible and practically deployable NLP systems.

8. Acknowledgements

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by

MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by *Junta de Andalucía*.

9. Bibliography

- Nesrine Bannour, Sahar Ghannay, Aurélie Névóol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools. In *Proceedings of the second workshop on simple and efficient natural language processing*, pages 11–21.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos. 2024. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599:128096.
- Benoit Courty, Victor Schmidt, Sasha Lucioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoieille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michał Stęchły, Christian Bauer, Lucas Otávio N. de Araújo, JPW, and MinervaBooks. 2024. mlco2/codecarbon: v2.4.1.
- Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.
- Daniel Hershovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in nlp research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494.

- Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527.
- Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.
- David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer.
- Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-del Arco, María Dolores Molina-González, María Teresa Martín-Valdivia, Luis Alfonso Ureña-López, and Arturo Montejó-Raéz. 2023. Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish. *Procesamiento del Lenguaje Natural*, 71:329–350.
- Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, and Arturo Montejó-Raéz. 2024. Environmental impact measurement in the mentalriskes evaluation campaign. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability@ LREC-COLING 2024*, pages 61–72.
- Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. 2021. Codecarbon: estimate and track carbon emissions from machine learning computing. *Cited on*, 20.
- Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.
- María Estrella Vallecillo-Rodríguez, María Victoria Cantero-Romero, Isabel Cabrera-de Castro, Luis Alfonso Ureña-López, Arturo Montejó-Raéz, and María Teresa Martín-Valdivia. 2024. Overview of refutes at iberlef 2024: Automatic generation of counter speech in spanish. *Procesamiento del Lenguaje Natural*, 73.
- Aimee Van Wynsberghe. 2021. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218.

10. Language Resource References

Álvarez-Ojeda, Pablo and Cantero-Romero, María Victoria and Semikozova, Anastasia and Montejó-Raéz, Arturo. 2025. *The precom-sm corpus: Gambling in spanish social media*.

Mármol Romero, Alba María and Moreno-Muñoz, Adrián and Plaza-Del-Arco, Flor Miriam and Molina-González, M. Dolores and Montejó-Raéz, Arturo. 2024. *MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish*. ELRA and ICCL.

Not Everything Is Greenwashing: Limitations of Automatic Analysis of Sustainability Reports, and a Proposal

Maria Pilar Uribe Silva, Rik van Noord, Malvina Nissim

CLCG, University of Groningen
{m.p.uribe.silva, m.nissim, r.i.k.van.noord}@rug.nl

Abstract

Sustainability reports (SRs) are essential for holding companies accountable, and they are required by law. They also serve as a key communication tool through which companies shape their image and disclose non-financial information. However, the rapid growth of these reports, their lack of standardisation, and the frequent use of strategically ambiguous language make it difficult for stakeholders to evaluate whether sustainability claims are genuine or deceptive. Previous work has focused on extracting misleading climate-related content and identifying greenwashing. We argue that this is not enough, because deception does not only appear in overtly false or misleading green claims, and it often emerges through a variety of subtle linguistic strategies. We therefore propose the development of a framework based on deception theories to examine how deceptive language operates in SRs, and we outline the challenges that should be taken as an invitation for future research.

Keywords: Deception, Deceptive Language, Sustainability Reports, Greenwashing

1. Introduction and Motivation

Addressing the climate and human rights crises has been recognised as a legal obligation by the Inter-American Court of Human Rights¹, not only for states but also for companies. Decisions such as this reinforce the responsibility of corporations to demonstrate transparency and accountability in their environmental and social actions. In this context, Sustainability Reports (SRs)² have become central – and mandatory by law – to how companies present their commitments and actions toward environmental and social goals. For companies, these reports are also valuable corporate assets (Dhir, 2005). Through them, they shape reputation, build trust among stakeholders (Tata and Prasad, 2014; Diouf and Boiral, 2017), and provide financial and non-financial information that supports and impacts investment decisions (Abideen and Fuling, 2024). On the other hand, for stakeholders, including capital markets, governments, and civil society, these documents are key tools for assessing whether corporate actions align with stated sustainability commitments (Ellimäki et al., 2023).

Over the past decade, both the number and diversity of SRs have grown considerably. According to a 2024 survey³, 96% of the world's 250 largest

companies now publish such reports. At the same time, the language of these documents has become increasingly complex. Research highlights issues of readability (Smeuninx et al., 2020; Pombinho et al., 2024), lexical consistency (Chalmers and Klingler-Vidra, 2023), and performance measurement (Wang et al., 2018). Despite the efforts of global frameworks such as the Global Reporting Initiative (GRI), the UN Sustainable Development Goals (SDGs), and the Sustainability Accounting Standards Board (SASB), there is still no mandated format or shared terminology, which leaves room for companies to adopt discourse strategies that appear responsible and positive – a well-known practice called *greenwashing* – while avoiding concrete accountability.

Parallel to the growth of SRs, research on misleading language in sustainability communication has also grown, following two main paths. One path has seen the surge of **manual** and corpus-assisted discourse studies, which have provided fine-grained insights into how deception is constructed linguistically: vagueness, exaggeration, and selective disclosure, for instance. These have contributed to revealing the complexity and persuasive tone of corporate communication (Christensen et al., 2013; Elving et al., 2015; Siano et al., 2017; Nwagbara and Belal, 2019; Parlakkaya and Stenka, 2024; Kochkina et al., 2024), and how language is actively used to create social realities, rather than simply reflecting them (Ashcraft et al., 2009).

The other path concerns the development of **automatic** detection methods, based on natural language processing (NLP) techniques, which have

¹https://www.corteidh.or.cr/docs/opiniones/seriea_32_en.pdf

²For simplicity, we will refer to Sustainability Reports (SRs) as a general term throughout this paper. These include Environmental, Social, and Governance (ESG) reports, Corporate Social Responsibility (CSR) reports, Integrated Reports (IR) and other related titles used by organisations to disclose their sustainability-related activities and performance.

³<https://kpmg.com/dk/en/esg/>

[survey-of-sustainability-reporting-2024.html](#)

contributed to scaling up the analysis of corporate disclosures. The majority of them, however, have focused primarily on detecting climate-related claims and greenwashing, which is too broad and unspecific to capture the complex and strategic nature of misleading corporate language. Moreover, they concentrate on the content of the statements rather than on their structure or form, where deception often lies. This limitation is evident in traditional NLP methods, which require clear, observable, and operational linguistic categories, rather than a polymorphic and dynamic term (Calamai et al., 2026; Gorovaia and Makrominas, 2025; Sneideriene and Legenzova, 2025; Stambach et al., 2023; Schimanski et al., 2024; Mohammadrezaei et al., 2024). Modern LLM-based approaches present other shortcomings: while they are able to detect explicit environmental claims and inconsistencies at scale, their performance drops when dealing with subtle, context-dependent strategies such as selective disclosure, vague commitments, and aspirational rhetoric (Velutharambath et al., 2026; Salminen et al., 2025). In addition, these models may align with prevalent misconceptions (Velutharambath et al., 2026). A statement can be factually correct and still be deceptive. For example, the sentence *"We reduced our carbon emissions by 20% compared to last year"* may be factually correct, but is deceptive if the reduction was due to the temporary closure of a facility, while overall long-term emissions continue to rise.

Contributions We believe the current NLP approaches to sustainability reporting analysis would be much more impactful by taking into account the wealth of research on deceptive language in communication, corporate communication, and linguistics. In this **short position paper**, we propose examining how deceptive language manifests in these types of documents and moving beyond the detection of greenwashing to the detection of broader categories of deceptive linguistic cues in SRs. We highlight the need for a conceptual framework grounded in deception theories and corporate strategies to guide their identification. This shift towards a transdisciplinary approach helps us see how deception operates in SRs, not only what these reports claim but also how they say it.

2. Related work and our position

The growing body of SRs highlights a clear need for more systematic and scalable evaluation. In this section, we discuss the foundational contribution of manual analysis techniques to detect linguistic cues of deception in the development of automated approaches. We then argue why this is insufficient and requires further advances, starting from a better framing of the problem.

Manual analysis Until now, a close reading analysis has played a fundamental role in uncovering corporate discourse strategies. As Siano et al. (2017) demonstrate in their content analysis of the Volkswagen emissions scandal⁴, deception in corporate discourse often takes the form of what they call "deceptive manipulation": a new type of irresponsible and deliberate corporate practice. Their work organised different kinds of greenwashing into a taxonomy that they then expanded, bringing two crucial insights: first, that contextual and critical reading helps uncover how companies use complex discursive strategies (Catenaccio et al., 2012; Painter and Martins, 2017) to construct a gap between discourse and action (Elving et al., 2015); and second, that breaking down this broad concept is essential to capturing the complexity of such deceptive practices. Unfortunately, the quantity and persuasive tone of these reports, combined with their qualitative structure and legal complexity, make them particularly difficult and time-consuming to analyse manually, even for professionals who have been working in the sector for a long time. For this reason, developing automatic methods to analyse these kinds of documents effectively and efficiently has become essential.

NLP methods Given the above, several studies have focused on using NLP methods to detect misleading practices, particularly greenwashing. Polignano et al. (2022) developed a system that uses NLP and information extraction methods to identify references to various sustainability topics within reports that comply with Global Reporting Initiative standards (GRI), improving the accessibility of these disclosures. Similarly, Gutierrez-Bustamante and Espinosa-Leal (2022) employed text mining techniques to evaluate the consistency of Nordic companies' sustainability reports with the GRI framework, demonstrating the advantage of models such as Latent Semantic Analysis (LSA) and Global Vectors for Word Representation (GloVe) in this context. Luccioni et al. (2020) presented ClimateQA⁵, a customised NLP model designed to identify climate-relevant sections through a question-answer approach, facilitating the extraction of relevant information. Gorovaia and Makrominas (2025) employed text analysis of corporate reports to identify inconsistencies in reporting between companies that had committed environmental violations and those that had not.

More recently, Calamai et al. (2026) reviewed 61 studies addressing one or more aspects of greenwashing detection in text. Their findings show that there are no datasets explicitly labelled for green-

⁴<https://www.epa.gov/vw/learn-about-volkswagen-violations>

⁵<https://www.climateqa.com/>

Page	Text excerpt	Deceptive	Greenwashing
17	We are ensuring that we have a well-balanced, healthy and relevant store portfolio in each market.	Yes	No
31	[...] improve the resilience of our business and supply chain, and empower our customers to be part of our industry's transformation	Yes	No
45	H&M Group is a customer-focused, creative, value-driven, responsible fashion and design company	Yes	Yes
53	Good procurement practices and close cooperation with suppliers are essential to ensuring that the company's products are always produced with consideration for people and the environment.	Yes	Yes
81	We are committed to growing our business in a sustainable way — by decoupling our growth from the use of finite natural resources	Yes	Yes
143	Our vision to lead the change means innovating, incubating and investing in scaling new materials.	Yes	No
143	Providing convenient ways for our customers to engage in circular fashion.	Yes	Yes

Table 1: Examples of deceptive statements from the H&M Sustainability Report (2023). Not all of them meet the definition of greenwashing in the European Union's Green Claims Directive Proposal.

washing (also due to the ambiguity of the term), and many current studies fragment global detection into intermediate tasks (e.g., detection of climate-related claims (Stambach et al., 2023), identification of statements that may indicate greenwashing (Bingler et al., 2024), and tone analysis (Kang and Kim, 2022; Polignano et al., 2022).

Undoubtedly, these studies have helped to extend the analysis beyond what manual approaches can do. But the fundamental problem remains unresolved: inconsistencies and ambiguities in the use of language in SRs remain a challenge in judging whether a public statement reflects a genuine commitment or not. Without a clear conceptual understanding of how deception works in corporate language, it is hard to tell the difference between strategic ambiguity and genuine uncertainty.

LLM-based tools The widespread use and continuous development of Large Language Models (LLMs) have brought them into focus as valuable tools in detecting misleading information in SRs. The ESGReveal system proposed by Zou et al. (2023) enhances LLMs with Retrieval-Augmented Generation (RAG) to retrieve structured ESG information, offering a benchmark for corporate reporting. Bronzini et al. (2024) used LLMs to construct knowledge graphs to analyse the disclosure of ESG in SRs. The work done by Usmanova and Usbeck (2024) proposes a conceptual framework for subdomain-specific pre-training of LLMs to improve the detection of green claims in corporate reports. And recently, Mishra et al. (2024) started offering an open-access and collaborative platform, DocQA⁶, for extracting information from technical and unstructured documents, such as SRs, through a conversational assistant that answers questions.

While the potential of this technology is clear, important challenges remain. These models can flag potential cases of greenwashing and even spot

certain linguistic cues of deception, but their analyses remain vague, show a tendency toward truth bias (Velutharambath et al., 2026; Markowitz and Hancock, 2023), and often lack adequate explanation. This vagueness does not only stem from the current technical limitations of LLMs, but it is also a product of imprecise language. Model performance is likely negatively affected by the use of existing definitions of greenwashing, which are at the same time overly broad and too limited to be effective in a highly domain-specific task such as SR analysis (Calamai et al., 2026). As a result, automated systems may flag the wrong signals or overlook more subtle forms of deception in SRs.

The problem Despite the valuable contributions of the aforementioned works, the lack of a clear, holistic framework that draws from multiple disciplines to identify deceptive discourse strategies within sustainability narratives at scale undermines the goal of analysing the fast-growing number and size of SR documents. This is where the core of the issue lies: **detecting deception in corporate sustainability disclosures requires more than extracting information to identify greenwashing**. As shown in Table 1, the definition of greenwashing according to the European Union's Green Claims Directive Proposal⁷ does not cover deceptive linguistic strategies that appear in sustainability-related publications. Vague or non-specific statements, as well as the omission of relevant information, can function as deceptive practices even when no explicit false claim is made. This shows that focusing only on explicit claims is not enough to capture how deception appears in these texts. Such complexity demands a deeper understanding of how language is used by companies to blur boundaries.

⁷See § 1.2: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2023:0166:FIN>

⁶<https://ds4sd.github.io/>

Our proposal Instead of focusing on greenwashing or the detection of climate-related claims, we propose to concentrate on deceptive corporate strategies. We argue that deceptive language does not always appear in overtly false or misleading green claims; rather, it emerges through a variety of subtle linguistic strategies that distort or obscure meaning. This issue has been extensively explored in the field of deception studies from a linguistic perspective, which provides valuable theoretical and methodological insights for developing a conceptual framework that makes the detection of deception in SRs more precise, operationalisable, and critically informed. By operationalising such a framework into automatic tools, we can move toward a more systematic and accurate analysis of deceptive language in SRs. As an initial step, this can be approached as a sentence and paragraph-level classification task focused on identifying observable deceptive discourse cues.

We draw on deception theory as a conceptual lens, but we do not assume that deceptive intent can be directly recovered from text alone. What can be operationalised are **observable linguistic and discourse phenomena**, such as vagueness, equivocation, obfuscation, and selective disclosure, which may function as deceptive strategies in a corporate context. We therefore distinguish between textual evidence and the stronger interpretive claim that such evidence reflects deceptive intent.

In the next section, we discuss the main theories of deception that underpin our position and explain the opportunity to create a conceptual framework grounded in deception theories, illustrating how deception manifests itself in corporate sustainability communication. In Section 4 we then review the main challenges ahead for operationalising and practically implementing such a framework.

3. Deception as Communicative and Linguistic Strategy

When we talk about deception in language, we talk about *intention*. The intention to hide information, to be ambiguous, and to avoid full cooperation in communication. But intention is not directly observable and cannot be reliably identified from language alone. Even so, research on deception has developed conceptual distinctions and methodological approaches that help scholars define deception, describe its main forms, and analyse how it may be realised in discourse.

We draw on these theories to examine what deception may look like in the context of SRs, how it may operate, and which linguistic and discourse-level cues may function as deceptive strategies in this setting. These include forms such as vagueness, equivocation, selective disclosure, and obfus-

cation. In this sense, deception theory serves as a conceptual lens, while the analysis itself focuses on observable textual evidence rather than on directly inferring hidden intent.

Theoretical foundation Galasinski (2000) encourages a shift away from the conventional understanding of deception as outright falsehoods. Instead, he conceptualises deceptive communication as a form of linguistic manipulation that operates within the limits of truth and falsehood. From this perspective, deception is not simply about lying but involves more complex and indirect forms of meaning-making. It can therefore be understood as a **spectrum** of possible manifestations rather than a single act, where *intention* plays a central role (DePaulo et al., 1996; Masip et al., 2004; Mahon, 2007, 2015).

Ekman (1996) identifies deception as a communicative act that involves managing information, behaviour, and emotional expression to influence others' perceptions. Masip et al. (2004) further stress that deception is not the result of misunderstanding or error but a deliberate act of fabricating or manipulating factual or emotional information for strategic purposes. Buller and Burgoon (1996) in their Interpersonal Deception Theory (IDT) emphasise that deception is an interactive process in which both deceiver and receiver dynamically adapt to one another, rather than a unidirectional act.

These perspectives are relevant because they frame deception as a strategic, interactional, and goal-oriented act. When applied to the communicative processes of SRs, companies can be viewed as the senders (the deceivers) and stakeholders as the receivers (the target), where the senders deliberately seek to shape perceptions and gain advantage through selective or distorted communication.

In this sense, deception in language can be viewed as a strategic form of message design and impression management (Buller and Burgoon, 1996; Ekman, 1996), and it can take many forms. It may include, for example, exaggerations, subtle lies (DePaulo et al., 1996); falsification, concealment and equivocation (Buller and Burgoon, 1996); misdirection, strategic ambiguity and deflection (Smith et al., 2010); and vagueness and evasiveness (Galasinski, 2000). As Vrij (2008) points out, these many forms of deception are especially difficult to define and even harder to detect, precisely because they are often subtle, context-sensitive, and adaptive. Moreover, we have to add that deceptive communication is determined by cultural norms, language structure and context (Polignano et al., 2022).

Linguistic insights and automatic detection

Developing automatic methods to detect and analyse deceptive content is surely challenging, but there is work from the NLP community that shows the potential of this line of research. For example, [Levitan et al. \(2018\)](#), who studied automatic deception detection, analysed a set of linguistic features in interview dialogues. Other research has examined how deception manifests through pragmatic and stylistic cues such as hedging, indirectness, overgeneralisation, and shifts in lexical choice or modality ([Hauch et al., 2015, 2016](#); [Zhou et al., 2004](#)). These studies show that deception can be traced through measurable linguistic signals, yet they also highlight the strong context dependence of such patterns.

This issue is discussed in depth in the work of [Gillings \(2024\)](#), who identified markers such as increased certainty or vagueness, and the strategic use of positive sentiment or abstract language in text. However, as Gillings notes, context remains a key limitation in detection. Recent studies further illustrate both the limits and potential of computational approaches: for instance, [Velutharambath et al. \(2026\)](#) demonstrate low cross-domain generalisability of linguistic cues, while [Salminen et al. \(2025\)](#) combine psycholinguistic features with transformer models to detect deceptive reviews.

While linguistic research has provided detailed qualitative insights into how deception operates through language, these findings have not yet been systematically translated into operational frameworks to understand how deceptive language operates in the context of sustainability. As mentioned, automatic systems still rely predominantly on surface-level features (e.g., word frequency, sentiment, syntactic patterns) and fail to integrate deeper discourse or pragmatic mechanisms. We believe that this gap limits their ability to capture context, intent, and communicative strategy, and to model the textual traces through which potentially deceptive discourse is realised.

4. Challenges

We are aware that, even with a strong framework and previous research supporting the conceptualisation of deception in SRs, our proposal still faces significant challenges when looking at the development of automated tools for analysing the deceptive discourse strategies used by companies. Here, we outline a few.

1. Lack of a domain-specific dataset To our knowledge, there is no dataset explicitly annotated for deceptive linguistic strategies in SRs, making it a necessity to create one to apply such a theoretical and conceptual framework. Importantly, our dataset will focus on observable linguistic and discourse cues rather than inferred authorial intent, and should define a clear unit of analysis, for example, at the sentence and paragraph level. The selection, curation, and maintenance of such a dataset is a collaborative and fundamental effort towards the creation of annotated data and, eventually, the training and evaluation of models that identify deceptive language in this context.

Deceptive discourse in corporate communications is often subtle, context-dependent and sector-specific. For this reason, annotation guidelines must clearly distinguish between observable textual evidence and broader interpretative claims about deception. This is important not only for consistency, but also to ensure the task is feasible. Furthermore, future evaluation should assess whether models generalise across different sectors, writing styles, and repeated text, rather than merely learning industry-specific patterns.

2. From PDF to text SRs are often published in highly variable, multimodal formats (graphs, images, infographics, etc.), usually as non-machine-readable PDFs.⁸ This makes it difficult to extract structured and comparable information ([Peng et al., 2024](#); [Mahadevkar et al., 2024](#); [Gupta et al., 2025](#)). Even if it were possible to extract clean content, the conversion from PDF to plain text poses the risk of losing information and relevant details, such as the position of elements and visual information, where companies also apply deceptive communication strategies.

3. LLM-written SRs LLMs already support many writing processes in the business and corporate sector ([Chiarello et al., 2024](#)), and companies will increasingly rely on these systems to produce content and documents such as SRs. This raises an additional challenge for future research. If LLMs begin to shape the language of SRs more strongly, the form and distribution of deceptive discourse cues may also change. Reports may become more standardised, polished, and stylistically homogeneous, which could make deceptive strategies harder to identify. This raises questions such as: how will the use of language, and especially deceptive language, change in these reports in the coming years? And how would this affect LLM-based methods applied in the detection of deceptive language?

⁸<https://www.responsibilityreports.com/>

4. Human judgment The role of human judgment is a central challenge. Our aim is to identify linguistic cues that may signal deceptive language in SRs and support their automatic analysis. Advances in LLM capabilities offer significant potential for analysing written texts, but these models can still misclassify, especially when dealing with subtle rhetorical and discursive strategies that demand strong contextual and pragmatic understanding (Kaya and Ghosh, 2024; Velutharambath et al., 2026). Even recent methodologies that improve LLM performance and evaluation (Kaya and Ghosh, 2024) do not fully resolve this issue. Human involvement, therefore, remains essential, not only for validating model outputs, but also for interpreting whether observable textual cues justify broader claims about deception. This is particularly important because such claims go beyond the text itself and depend on context, background knowledge, and pragmatic judgment.

5. Conclusion

Research on how to detect deception has always advanced through collaboration and interdisciplinarity. The increasing quantity and complexity of SRs urgently demand these efforts to be renewed through a more scalable and flexible approach. In this short position paper, we respond to that challenge by proposing a broader conceptual framework grounded in deception theories, with the potential to enable more accurate, linguistically informed automatic detection of deceptive language. Rather than focusing only on explicit green claims, we argue for attention to observable discourse cues that may signal deceptive communication in SRs.

Our aim is to understand how deceptive language operates within SRs on a large scale, from linguistic, pragmatic and discourse-analytical perspectives, while keeping the focus on textual evidence rather than inferred intent. Building on previous efforts in automatic deception detection, we invite the NLP community to take a critical discourse perspective on how companies use language in their public-facing statements about sustainability, and how they may shape stakeholders' perceptions through textual strategies such as vague commitments, obfuscation, and selective disclosure, especially when these are difficult to verify through surface-level analysis alone.

6. Bibliographical References

- Zain Abideen and Han Fuling. 2024. Sustainability reporting and investor sentiment. A sustainable development approach to Chinese-listed firms. *Journal of Cleaner Production*, 466:142880.
- Karen Lee Ashcraft, Timothy R. Kuhn, and François Cooren. 2009. 1 constitutional amendments: “materializing” organizational communication. *Academy of Management Annals*, 3(1):1–64.
- Julia Anna Bingler, Mathias Kraus, Markus Leipold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.
- Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? Deriving structured insights from sustainability reports via large language models. *EPJ Data Science*, 13.
- David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory*, 6(3):203–242.
- Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2026. Detecting greenwashing: A natural language processing literature survey.
- Paola Catenaccio et al. 2012. *Understanding CSR discourse: Insights from linguistics and discourse analysis*. Arcipelago edizioni.
- Adam Chalmers and Robyn Klingler-Vidra. 2023. Inconsistent language can undermine the credibility of corporate social responsibility reports.
- Filippo Chiarello, Vito Giordano, Irene Spada, Simone Barandoni, and Gualtiero Fantoni. 2024. Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, 133:103002.
- Lars Thøger Christensen, Mette Morsing, and Ole Thyssen. 2013. CSR as aspirational talk. *Organization*, 20(3):372–393.
- Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. 1996. Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5):979.
- Krishna S. Dhir. 2005. The value of language: concept, perspectives, and policies. *Corporate Communications: An International Journal*, 10(4):358–382.

- Dominique Diouf and Olivier Boiral. 2017. The quality of sustainability reports and impression management: A stakeholder perspective. *Accounting, Auditing & Accountability Journal*, 30(3):643–667.
- Paul Ekman. 1996. *Why don't we catch liars?* *Social Research*, 63(3):801–817.
- Pia Ellimäki, Ruth Aguilera, Nuria Hurtado-Torres, and J. Aragon-Correa. 2023. *The link between foreign institutional owners and multinational enterprises' environmental outcomes.* *Journal of International Business Studies*, 54:910–927.
- Wim Elving, Ursa Golob, Klement Podnar, Anne Ellerup Nielsen, and Christa Thomsen. 2015. *Guest editorial: The bad, the ugly and the good: New challenges for CSR communication.* *Corporate Communications An International Journal*, 20:118–127.
- Dariusz Galasinski. 2000. *The language of deception: A discourse analytical study.* Sage Publications.
- Mathew Gillings. 2024. *Corpus Linguistic Approaches to Deception Detection.* Routledge.
- Nina Gorovaia and Michalis Makrominas. 2025. *Identifying greenwashing in corporate-social responsibility reports using natural-language processing.* *European Financial Management*, 31(1):427–462.
- Tanay Gupta, Tushar Goel, and Ishan Verma. 2025. *Exploring multimodal language models for sustainability disclosure extraction: A comparative study.* In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 141–149, Albuquerque, New Mexico. Association for Computational Linguistics.
- Marcelo Gutierrez-Bustamante and Leonardo Espinosa-Leal. 2022. *Natural language processing methods for scoring sustainability reports—a study of Nordic listed companies.* *Sustainability*, 14(15).
- Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. *Are computers effective lie detectors? A meta-analysis of linguistic cues to deception.* *Personality and Social Psychology Review*, 19(4):307–342. PMID: 25387767.
- Valerie Hauch, Siegfried L. Sporer, Stephen W. Michael, and Christian A. Meissner. 2016. *Does training improve the detection of deception? a meta-analysis.* *Communication Research*, 43(3):283–343.
- Hyewon Kang and Jinho Kim. 2022. *Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods.* *Applied Sciences*, 12(11).
- Zeyneb N. Kaya and Souvick Ghosh. 2024. *Decoding large-language models: A systematic overview of socio-technical impacts, constraints, and emerging questions.*
- Nataliia Kochkina, Silvia Macchia, and Michela Floris. 2024. *Strategic language use in sustainability reporting: An empirical study.* *Sustainability*, 16(23).
- Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. 2018. *Linguistic cues to deception and perceived deception in interview dialogues.* In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, Louisiana. Association for Computational Linguistics.
- Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. *Analyzing sustainability reports using natural language processing.*
- Supriya Mahadevkar, Shruti Patil, Ketan Kotecha, Lim Soong, and Tanupriya Choudhury. 2024. *Exploring ai-driven approaches for unstructured document analysis and future horizons.* *Journal of Big Data*, 11.
- James Mahon. 2007. *A definition of deceiving.* *International Journal of Applied Philosophy*, 21:181–194.
- James Mahon. 2015. *The definition of lying and deception.* *Stanford Encyclopedia of Philosophy.*
- David Markowitz and Jeffrey Hancock. 2023. *Generative AI are more truth-biased than humans: A replication and extension of core truth-default theory principles.*
- Jaume Masip, Eugenio Garrido, and Carmen Herero. 2004. *Defining deception.* *Anales de Psicología*, 20:147–171.
- Lokesh Mishra, Cesar Berrospi, Kasper Dinkla, Diego Antognini, Francesco Fusco, Benedikt Bothur, Maksym Lysak, Nikolaos Livathinos, Ahmed Nassar, Panagiotis Vagenas, Lucas Morin, Christoph Auer, Michele Dolfi, and Peter Staar. 2024. *ESG accountability made easy: DocQA at your service.* *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23814–23816.

- Mahsa Mohammadrezaei, José Carlos Marques, and Asif Huq. 2024. Use of text mining and natural language processing techniques in analyzing sustainability reports: A systematic literature review and assessment.
- Uzoechi Nwagbara and Aatur Belal. 2019. Persuasive language of responsible organisation? a critical discourse analysis of corporate social responsibility (CSR) reports of Nigerian oil companies. *Accounting, Auditing & Accountability Journal*, 32(8):2395–2420.
- Christopher Painter and Jorge Martins. 2017. Organisational communication management during the Volkswagen diesel emissions scandal: A hermeneutic study in attribution, crisis management, and information orientation. *Knowledge and Process Management (Print Edition): the journal of corporate transformation*, 2017.
- Emre Parlakkaya and Renata Stenka. 2024. *Analysing discourse in corporate sustainability reporting: linguistic insights*, pages 456–473.
- Jiahui Peng, Jing Gao, Xin Tong, Jing Guo, Hang Yang, Jianchuan Qi, Ruiqiao Li, Nan Li, and Ming Xu. 2024. Advanced unstructured data processing for ESG reports: A methodology for structured transformation and enhanced analysis.
- Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An NLP approach for the analysis of global reporting initiative indexes from corporate sustainability reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–8, Marseille, France. European Language Resources Association.
- Miguel Pombinho, Ana Fialho, and Jorge Novas. 2024. Readability of sustainability reports: A bibliometric analysis and systematic literature review. *Sustainability*, 16(1).
- Joni Salminen, Mekhail Mustak, Soon-Gyo Jung, Hannu Makkonen, and Jim Jansen. 2025. Decoding deception in the online marketplace: Enhancing fake review detection with psycholinguistics and transformer models. *Journal of Marketing Analytics*, pages 1–18.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.
- Alfonso Siano, Agostino Vollero, Francesca Conte, and Sara Amabile. 2017. “More than words”: Expanding the taxonomy of greenwashing after the Volkswagen scandal. *Journal of Business Research*, 71:27–37.
- Nils Smeuninx, Bernard De Clerck, and Walter Aerts. 2020. Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and NLP. *International Journal of Business Communication*, 57(1):52–85.
- Sandi W Smith, Steven R Wilson, Judee K Burgoon, and Timothy R Levine. 2010. Advances in deception detection. In *New Directions in Interpersonal Communication Research*, pages 201–220. SAGE Publications, Inc.
- Agne Sneideriene and Renata Legenzova. 2025. Greenwashing prevention in environmental, social, and governance (ESG) disclosures: A bibliometric analysis. *Research in International Business and Finance*, 74:102720.
- Dominik Stambach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.
- Jasmine Tata and Sameer Prasad. 2014. CSR communication: An impression management perspective. *Journal of Business Ethics*, 132.
- Aida Usmanova and Ricardo Usbeck. 2024. Structuring sustainability reports for environmental standards with LLMs guided by ontology. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 168–177, Bangkok, Thailand. Association for Computational Linguistics.
- Aswathy Velutharambath, Kai Sassenberg, and Roman Klinger. 2026. What if deception cannot be detected? A cross-linguistic study on the limits of deception detection from text.
- Aldert Vrij. 2008. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.
- Xuepeng Wang, Kang Liu, and Jun Zhao. 2018. Detecting deceptive review spam via attention-based neural networks. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 866–876. Springer.

Lina Zhou, Judee K. Burgon, Douglas P. Twitchell, Tiantian Qin, and Jay F. Nunamaker Jr. 2004. [A comparison of classification methods for predicting deception in computer-mediated communication](#). *Journal of Management Information Systems*, 20(4):139–166.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, Zongxiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2023. [ESGReveal: An LLM-based approach for extracting structured data from ESG reports](#). *ArXiv*, abs/2312.17264.

Author Index

- Bardelli, Alessandro Pietro Bardelli, 79
Basili, Roberto, 15
Batista-Navarro, Riza, 42
Bosco, Cristina, 87
Brito, Mariana Madruga de, 115
- Calabrese, Armando, 15
Chierchiello, Elisa, 87
Christopoulou, Aikaterini, 105
Costa, Roberta, 15
Croce, Danilo, 15
- Demirtaş, İrem, 79
Detsis, Vassilis, 105
Di Bari, Marilena, 79
Di Palma, Eliana, 87
- Ella, Yanni Jose C., 42
- Filannino, Michele, 79
- Gabud, Roselyn Santos, 42
Gatos, Basilis, 105
Glazkova, Anna, 135
- Jakob, Charlott, 1
- Lastimoso, John Michael L., 42
Laviste, Monica Ashley R., 42
Lebedeva, Daria, 135
Lord, Sabina Fiebig, 60
- Madureira, Brielen, 115
Mármol-Romero, Alba María, 144
Möller, Sebastian, 1
Montejo-Raez, Arturo, 144
Moreno Muñoz, Adrián, 144
Mousavian Anaraki, Seyed Alireza, 15
- Nguyen, Huyen Phuong, 69
Niekler, Andreas, 115
Nissim, Malvina, 154
- Pannitto, Ludovica, 87
Prange, Jakob, 26
- Raspanti, Federico, 79
- Roccia, Mariana, 60
Rukh, Mah, 60
- Sahitaj, Premtim, 1
Santiañez, Wilfred John E., 42
Scala, Simona, 79
Schmitt, Vera, 1
Schüßler, Benjamin Josef, 26
Solopova, Veronika, 1
Sophocleous, Andry, 53
- Tiburzi, Luigi, 15
Tiernan, Eleanor, 60
- Upravitelev, Max, 1
Uribe Silva, Maria Pilar, 154
- Vallego, Jorge, 60
van Noord, Rik, 154
- Zakharova, Olga, 135