ClimateNLP 2026

The 3rd Workshop on Natural Language Processing Meets Climate Change

Proceedings of the Workshop

October 28, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-465-1

Introduction

We are delighted to welcome you to ClimateNLP 2026, the third Workshop on Natural Language Processing Meets Climate Change. This year, the workshop is co-located with EMNLP 2026 and takes place on October 28, 2026, in the beautiful city of Budapest, Hungary.

ClimateNLP aims to be the premier publication venue for research at the intersection of Natural Language Processing (NLP) and climate change. The workshop brings together researchers and practitioners to discuss how NLP methods can be incorporated into climate change science and climate change action. This year, the program includes four keynote talks by Chung-Chi Chen (National Institute of Informatics), Ramit Debnath (University of Cambridge), Maria João Sousa (Climate Change AI), and Nelly Bencomo (Durham University). Furthermore, we hold a panel discussion on future directions for ClimateNLP. Four oral paper presentations and two poster sessions round up the day.

We received 37 submissions this year and recruited 38 active Program Committee (PC) members who are experts in NLP, climate change, or both. The vast majority of submissions received two reviews. When making our selections for the program, we carefully considered the reviews and held extensive discussions among the members of the organizing committee. The members of the Program Committee did an excellent job in reviewing the submitted papers, and we thank them for their essential role in selecting the accepted papers and helping produce a high-quality program for the workshop. In line with our goal of discussing and learning about the intersection of NLP and climate change, we aimed to create an inclusive program that accommodates as many favorably rated papers as possible. We accepted 24 papers (acceptance rate 64.9%), 21 of which are included in these proceedings. The accepted papers span a broad range of topics, including climate-related benchmarks and datasets, information extraction from corporate sustainability and ESG disclosures, analysis of climate discourse in news and social media, climate adaptation, and the efficiency and environmental footprint of language models themselves.

We thank our Program Committee members for committing their time to help us select an excellent technical program.

We thank all the authors who submitted to the workshop and all workshop participants for making ClimateNLP 2026 a success and for growing the research area of NLP for climate change with their fine work.

The ClimateNLP 2026 Organizing Committee

Organizing Committee

Program Chairs

Tobias Schimanski, University of Zurich, Switzerland
Aida Usmanova, Leuphana University Lüneburg, Germany
Gaku Morio, Hitachi, Ltd. and Hitotsubashi University, Japan
Jingwei Ni, ETH Zurich, Switzerland
Kalyan Dutia, Climate Policy Radar, UK
Seyed Saeid Vaghefi, University of Zurich, Switzerland
Dominik Stammbach, Princeton University, USA
Markus Leippold, University of Zurich, Switzerland

Program Committee

Reviewers

Shafiuddin Rehan Ahmed, Ekant Muljibhai Amin

Nan Bai, Jetske Bonenkamp, Marco Bronzini

Henry Franks

Nupoor Gandhi, Dario Garigliotti, Francesca Grasso

Qi He, Angel Hsu

Aditya Jain, Eshaan Jain

Nikhil Avinash Kharwandikar, Sai Koneru, Mogana Kumaran, Murathan Kurfali

Kai-Robin Lange, Valentin Lemaire, Yucheng Lu

Diego Manya

Yusuke Ohtsubo

Srijith Ravikumar, Tim Repke, Jonas Rieger, Harri Rowlands

Anna Steinberg Schulten, Anne J Sietsma, Manfred Stede

Tommy Tran, Mark Tyrrell

Tejesh Vaish

Azmine Toushik Wasi, Hendrik Weichel

Lexing Xie

Zukang Yang

Shorouq Zahra, Ji Zhang

Table of Contents

How Loud Rumbles Hit Newsstands: A Data Analysis of Coverage and Spatial Bias in German News about Landslides Around the World
Brielen Madureira, Andreas Niekler, Marc Keuschnigg and Mariana Madruga de Brito 1

CA-ClimateNews: A Human-Annotated Multilingual Benchmark for Climate-Relevance Classification of Central Asian News
Adilet Metinov, Zarina Uvalieva, Bektemir Kumarbai Uulu, Nurtilek Alibekov and Tynchtykbek Tashbaltaev 20

Climatextract: A Systematic Prompt Evaluation For Climate Information Extraction
Anna Steinberg Schulten, Malte Schierholz, Lisa Reichenbach, Frauke Kreuter and Alexander Fraser 29

Who Warmed the Archives? LLMs Overestimate Historical Warmth
Claudiu Creanga and Liviu P Dinu 41

Improving Efficiency of Ternary Large Language Models through Pruning
Pietro Tropeano, Maria Maistro, Tuukka Ruotsalo, Christina Lioma and Nikolaos Aletras 51

Does Tropical Adaptation Help Plant Language Models Generalize? A Split-Sensitivity Evaluation
Gabriela Morales Soto and Helio Pedrini 85

Extracting information on public water shortages using news media and LLMs: a pilot study in Sweden
Jeanne Fernandez, Shorouq Zahra, Mariana Madruga de Brito and Johanna Mård 99

Ontology-Guided Query Expansion for Evidence Retrieval in ESRS E1 Climate Disclosures
Tatsuya Ishisaka 116

The Climate Contracting Knowledge Graph: Construction and Evaluation
Georgia Ray and Felix Cohen 124

Beyond Dictionaries: Machine-Learned Normalization of Emission Units in Information Extraction
Lisa Reichenbach, Simon Oehler, Malte Schierholz and Anna Steinberg Schulten 132

FoodClimNER: A Manually Annotated Corpus for Food Practice Entities Related to Climate Change
Mathieu Laï-king, Arnaud Ferré, Louise Deléger, Yann Le Bodo and Estelle Chaix 143

Fine-Tuned Encoders versus Zero-Shot Prompting for Detecting Threat and Solution Framing in German Climate News
Raven Adam, Marie Lisa Kogler and David Maier 158

Retrieving Climate Change Adaptation Concepts from World Heritage Management Plans in the Netherlands
Vedant Jumle, Orhan Agaoglu, Laurens Mol, Edgar Salas Gironés and Nan Bai 170

ESG-CID-Plus: A Stratified Benchmark for Disclosure Content Index Table Extraction from Corporate Sustainability Reports
Shafiuddin Rehan Ahmed, Reza Abdi, Tianyu Luan, Teja Kanchinadam, Christopher Clarke and Eugene Siow 183

Evidence-Grounded Auditing of Identification Assumptions in Climate-Policy Causal Evaluations
Yonghong Zhang, Yong Xie, Isabel M. Parra and Ricardo Correia 199

The Energy Footprint of LLM-Based Environmental Analysis: LLMs and Domain Products
Alicia Bao, Angel Hsu, Diego Manya, Ji Zhang and Jiamian He 220

Can Multimodal Large Language Models Identify Thematic Maps in Scientific Articles?
Tomas Turek, Vincent Nguyen, Dhirendra Singh, Sarvnaz Karimi and Xiuzhen Zhang 231

AquaCrisis: A Multilingual Benchmark for Classifying Water-Utility Public Communication
Vishal Kumar and Mourad Oussalah 246

Climate and Migration on German Political Party YouTube: A Corpus with Stance, Propaganda, and Divisiveness Annotation
Zaher Alkaei, Manfred Stede and Christoph Maximilian Abels 282

Can Vision-Language Models Assess Environmental Reporting Compliance? A Benchmark for Requirement-Level Disclosure Alignment
Vaishali Aggarwal, Ishan Verma, Tanay Kumar Gupta and Tirthankar Dasgupta 298

Proportionate Responses: The Case for Small Language Models as Sustainable Expert Climate Policy Annotators
Jetske Bonenkamp, Robbert Biesbroek and Ioannis N. Athanasiadis 315

Program

Wednesday, October 28, 2026

09:00 - 09:05 *Opening Remarks*

09:05 - 09:30 *Keynote Speech 1: Beyond 2030: Turning Sustainability Commitments into Verifiable and Achievable Outcomes (Chung-Chi Chen)*

09:30 - 09:45 *Oral Presentation 1*

Improving Efficiency of Ternary Large Language Models through Pruning
Pietro Tropeano, Maria Maistro, Tuukka Ruotsalo, Christina Lioma and Nikolaos Aletras

09:45 - 10:30 *Poster Session 1*

10:30 - 11:00 *Coffee Break*

11:00 - 11:05 *Session 2 Introduction*

11:05 - 11:30 *Keynote Speech 2: Quantifying Climate Narratives, Public Sentiment, and Policy Reception via Computational Social Science (Ramit Debnath)*

11:30 - 11:45 *Oral Presentation 2*

Ontology-Guided Query Expansion for Evidence Retrieval in ESRS E1 Climate Disclosures
Tatsuya Ishisaka

11:45 - 12:30 *Panel Discussion - Future Directions for ClimateNLP with Chung-Chi Chen, Ramit Debnath, Maria João Sousa*

12:30 - 14:00 *Lunch Break*

14:00 - 14:05 *Session 3 Introduction*

14:05 - 14:30 *Keynote Speech 3: Public Interest Technology and Climate Change AI (Maria João Sousa)*

14:30 - 14:45 *Oral Presentation 3*

Wednesday, October 28, 2026 (continued)

Retrieving Climate Change Adaptation Concepts from World Heritage Management Plans in the Netherlands
Vedant Jumle, Orhan Agaoglu, Laurens Mol, Edgar Salas Gironés and Nan Bai

14:45 - 15:30 *Poster Session 2*

15:30 - 16:00 *Coffee Break*

16:00 - 16:05 *Session 4 Introduction*

16:05 - 16:30 *Keynote Speech 4: Beyond Accuracy: Bridging the Reasoning Gap in AI-Assisted Climate Law (Nelly Bencomo)*

16:30 - 16:45 *Oral Presentation 4*

Proportionate Responses: The Case for Small Language Models as Sustainable Expert Climate Policy Annotators
Jetske Bonenkamp, Robbert Biesbroek and Ioannis N. Athanasiadis

16:45 - 17:00 *Closing Remarks by Markus Leippold*

How Loud Rumbles Hit Newsstands: A Data Analysis of Coverage and Spatial Bias in German News about Landslides Around the World

Brielen Madureira[1,2], **Andreas Niekler**[1,3], **Marc Keuschnigg**[4,5], **Mariana Madruga de Brito**[2]

[1]LeipzigLab – Climate Discourse, Leipzig University, Germany
[2]Helmholtz Centre for Environmental Research, Germany
[3]Computational Humanities, Leipzig University, Germany
[4]Institute of Sociology, Leipzig University, Germany
[5]Institute for Analytical Sociology, Linköping University, Sweden

Correspondence: brielen.madureira@uni-leipzig.de

Abstract

Landslides often hit newsstands due to their destructive and potentially fatal effects. News are a valuable source of information for creating or enriching disaster databases and for expediting media-based studies of the dynamics of media attention. To accomplish that, news datasets must be filtered, geolocated and validated. This paper focuses on how landslides around the world are reported in German newspapers. We analyse almost 55k news articles about 4.5k news events in a 25-year period, compare it with external measures of countries' susceptibility to landslides and provide insights, e.g. the overreporting of Southern and Western Europe, to foster further studies on inequalities in media attention to international disasters.

1 Introduction

Sounds of cracking or breaking wood, groaning of the ground, a loud rumble and the sensation of a passing freight train. Poetic as they may sound, these words are used by the U.S. Geological Survey[1] to describe the warning signs of a hazardous type of event: landslides. Also known as mass movements, they are "the downslope movement of soil, rock, and organic materials under the effects of gravity", e.g. rockfalls and debris flow, triggered by water, seismic, volcanic or human activity (Highland and Bobrowsky, 2008) and also influenced by climate change (Gariano and Guzzetti, 2022). They can cause vast socioeconomic impact and fatalities, demanding mitigation and adaptation measures (Kjekstad and Highland, 2009; Petley, 2012).

Historical data on the occurrence and impact of landslides and other disasters worldwide are available in dedicated inventories such as EM-DAT (Delforge et al., 2025) and the Global Landslide Catalog (GLC) (Kirschbaum et al., 2009). However, global databases exhibit spatial biases, limited temporal coverage, missing data and specific inclusion criteria that lead to incompleteness (Gall et al., 2009; Jones et al., 2023). For instance, GLC contains only rainfall-triggered landslides and EM-DAT has strict inclusion thresholds (e.g. at least 10 fatalities) and relies on limited data sources.

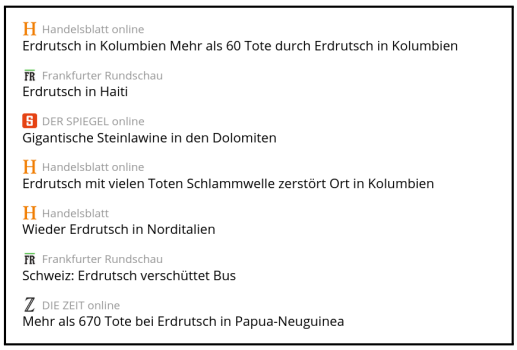


Figure 1: Headlines about landslides around the world in German news from the `wiso-net` database.

News have emerged as an alternative source of disaster information (see Section 2) but they stem from latent criteria (Galtung and Ruge, 1965; Östgaard, 1965; Joye, 2010; Alipour et al., 2024), such as the affected country's proximity or narrative appeal. Disproportions in media attention to landslides arise from the interplay of at least two factors: how often such disasters occur in reality and the newsworthiness of each event and country, given implicit media choices.

Datasets of news about disasters thus serve two purposes: (i) enriching existing disaster databases within a text-as-data paradigm for climate impact and adaptation research and (ii) unravelling inequalities in media attention to disasters in different places. Taking Germany as a case study, this paper provides an analysis of a dataset of geolocated news articles in German about landslides abroad (like those in Figure 1), its comparison to external measures, and preliminary insights into variations in media coverage of events across countries.

[1] www.usgs.gov/programs/landslide-hazards/what-are-signs-landslide-development-what-do-i-do-if-a-landslide-occurs

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 1–19
October 28, 2026 ©2026 Association for Computational Linguistics

2 Related Work

This work is framed by three key perspectives: efforts towards enhancing global landslide databases, news as a source of text-based disaster information, and studies of media attention to disasters.

Landslide databases Landslides are inherently context-specific phenomena shaped by local conditions. Information about their occurrence is often fragmented in various local sources with limited interoperability (Van Den Eeckhaut et al., 2013). As a result, studies often focus on single events (i.e. landslides in specific regions and time) or types (e.g. only rainfall-induced landslides). Indeed, studies often rely on national or sub-national databases (Mirus et al., 2020; Martha et al., 2021; Egas et al., 2024; Niyokwiringirwa et al., 2024; Bonini et al., 2025; Massey et al., 2025; Xie et al., 2026, *inter alia*). Fewer global databases are available to support systematic spatio-temporal analyses across regions and scales, as gathering and unifying data from various sources is a strenuous task due to their "different criteria, description details, and temporal and spatial scales" (Gómez et al., 2023). Besides, existing inventories can be biased toward areas with higher reporting capacity (Emberson et al., 2020).

Texts in climate impact and adaptation News archives and other collections of digital documents are a means to expand landslide databases with more events or metadata (Domínguez Cuesta et al., 1999; Taylor et al., 2015; Kreuzer and Damm, 2020; Franceschini et al., 2022). This is a particular case of a current trend in text-based climate impact research, aiming to create inventories of hazards and disasters based on information extraction from news texts (Sodoge et al., 2023; Alencar et al., 2024; Nguyen et al., 2024; Avcıoğlu et al., 2025; Valkenborg et al., 2026). News can additionally be used to validate models of landslides and other hazards such as risk assessment and warning systems (Battistini et al., 2017; Yagoub et al., 2020).

News geolocation and analysis The dynamics of how climate topics and events are covered in the news has been under active investigation (Yan and Bissell, 2015; Cai et al., 2025; Roth et al., 2025; Kong and Purves, 2026). We build upon those works from a new angle: how German newspapers portray landslides abroad. To geolocate news, we follow the practice of information extraction with Large Language Models (LLM) (Li et al., 2024; Bauman et al., 2025; Kriesch and Losacker, 2025).

3 Data Preparation

We derived a sample of news about landslides geolocated at the country level from the news dataset about extreme climate events and other disasters created by Madureira et al. (2026a). Details on the selection and annotation process can be found in that publication; here, only the relevant aspects are summarised.

A collection of 317,036 news articles in German from 2000 to 2024 was retrieved from the `wiso-net` news aggregator database[2] using 17 German landslide-related keywords (such as *landslide*, *rockfall* and *mass movement*; see Appendix). After the preprocessing steps to reduce the number of unrelated or local news, 88,940 candidate documents remained. Among them, 450 unique texts were randomly sampled (uniformly over the years), and annotated by humans with a binary label for relevance (i.e. whether the news refer to actual landslides) and the country where the landslide occurred. 58.6% of the annotated documents were relevant.

Processing We used the open-weight, locally run LLM `google/gemma-4-31b` to first identify all documents that report on specific past landslides and, then, geolocate (at the country level) the ones classified as relevant. This was performed in a two-pass procedure using the prompts displayed in the Appendix. In the first step, the model predicted a binary relevance label; in the second, it was prompted to output a JSON-formatted string with the ISO-3 code of the countries where landslides were reported.

3.1 Evaluation on the Gold Standard

To ensure the quality of the resulting dataset, we first assessed the LLM's performance on the human-annotated sample (i.e. the gold standard).

Relevance classification The model achieved binary precision, recall and F1-score of 0.865, 0.973 and 0.916, respectively. We manually inspected all 40 false positives. While 5 were judged as really not relevant, the remaining ones were either wrongly annotated or at least related to the overall *topic* of landslides, e.g. referring to the risk, possible impact or regularity of such events. Although we focus on landslides *de facto*, the inclusion of some false positives of this type is acceptable, since established risks can already lead to consequences such as official warnings or prevention measures.

[2] Available at `https://www.wiso-net.de/`.

Geolocation The accuracy was 0.875 for exact matches (i.e. the extracted countries fully matched the annotation) and 0.924 for overlaps (i.e. at least one extracted country matched the annotation). In the gold standard, 242 documents were about a single country, 19 about 2, and 3 about 3 countries. The LLM extraction followed a similar distribution: 241 single-country documents, 18 about 2 and 5 about 3. There was thus no evidence of long lists of spurious countries inflating the overlap accuracy.

These results indicate that the news sample had sufficient precision and recall and that the geolocation was accurate enough for subsequent analyses.

3.2 Independent External Datasets

The list of countries by the United Nations Statistics Division was used as a reference. After a few needed adjustments (see Appendix), 244 countries remained. For each country, we compiled the following country-level indicators to aid the analysis and validation (download links in the Appendix):

- **Landslide occurrence (EM-DAT)**: the number of landslide events (main or associated type) from 2000 to 2024 recorded in the International Disaster Database (2,071 in total).
- **Landslide frequency (WB-GLHM)**: the estimated average annual number of (flood and earthquake-triggered) landslides predicted by the World Bank's Global Landslide Hazard Map.
- **Landslide risk**: the landslide risk, categorised as high, medium, low and very low, computed by the ThinkHazard! project of the World Bank Group and the Global Facility for Disaster Reduction and Recovery.
- **Development status**: the country's development level (developed/developing), mapped to Global North and South, respectively, as disclosed by the United Nations Conference on Trade and Development.
- **Income group**: the income level (low, lower middle, upper middle or high) as classified by the World Bank.

4 Methods

We created a *count time series* (Davis et al., 2021) for each country, i.e. a series of non-negative integers corresponding to the daily number of news about landslides in the period from Jan 1, 2000 to Dec 31, 2024. Duplicate texts published in different news outlets were counted as distinct documents as they increase the level of media attention. Since some documents refer to landslide events in more than one country, separate *data instances* were created for each mentioned country.

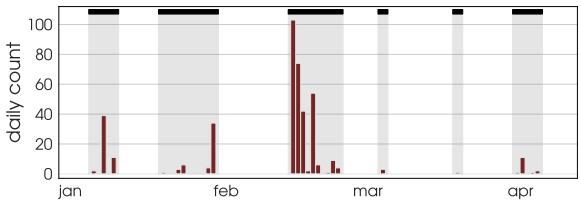


Figure 2: Example of the segmentation of a count time series into news events using the data instances about Brazil from Jan to mid-Apr, 2022. Grey-shaded areas represent each distinct identified news event.

Two steps followed: (i) the identification of *news events* and (ii) a comparison of the number of identified news events to external measures, in order to assess whether the distribution of landslide-related news aligned with expected landslide occurrence.

4.1 Identification of News Events

Each country's time series was segmented into *news events*, defined here as temporally contiguous clusters of news discussing landslides in that country. The number of active days (i.e. days with at least one document, representing non-zero media attention) *per country* is very sparse (see Section 5). We thus used a straightforward segmentation method: each sequence of adjacent active days was considered a distinct news event. A maximum of 4 sequential inactive days between active days was allowed during a news event. If 5 or more sequential inactive days occurred, news events were split. Figure 2 shows an example of the segmentation.

We measured the total number of documents, events and active days per country. For each identified news event, we computed its duration in days, its total volume of news documents and other measures listed in Table 6 (Appendix).

Note that not every news event aligns with *de facto* new landslides in the real world: while many texts do discuss ongoing or latest landslides, others recall past events, refer to aggregated observations or discuss the hazard without mentioning a concrete event. The distinction is left for future work; in this study, what matters is that each news event brings the *topic* of landslides in a country (back) into the media's attention.

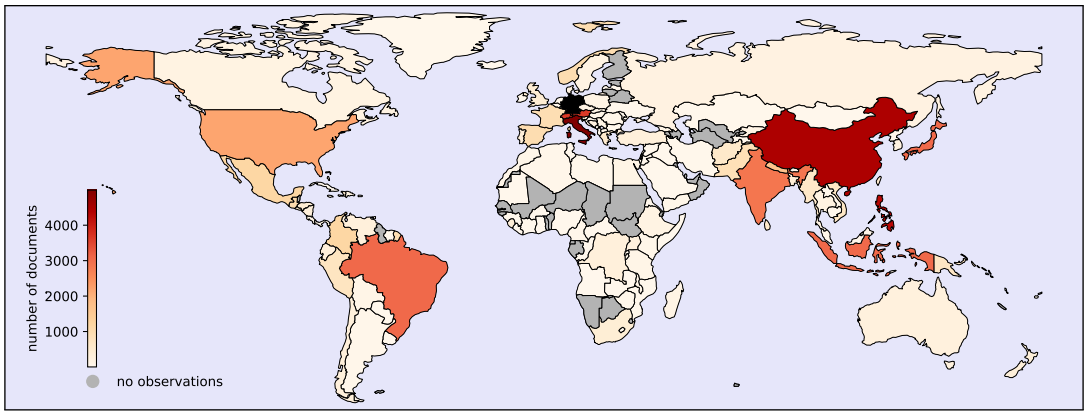


Figure 3: Total number of news documents about landslides in each country. Germany (in black) was not analysed.

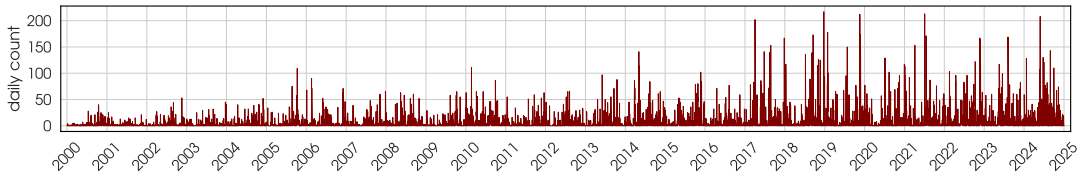


Figure 4: Overview of the count time series of daily news about landslides abroad in German newspapers. The increase in recent years may partially reflect the underlying temporal bias of the news database rather than necessarily an increase in landslide events or in media attention.

4.2 Salience Scores and Divergence

In line with the extra and intra-media data investigation outlined by Rosengren (1970), we verified to what extent our measures aligned with independent observations. The number of identified news events per country was compared with two external measures: the number of landslide events recorded in EM-DAT for the same period and the estimated annual average in WB-GLHM. This validation step is meant for evaluating whether countries receiving more media attention are the same with higher documented landslide occurrence or susceptibility, and for investigating potential biases in news coverage.

Since no authoritative and complete global database of *all* landslides exists, and news events do not always refer to new landslides, we refrained from comparing exact numbers. We were mainly interested in how many more landslides (or news events) occur or are expected in a country in comparison to other countries. For that purpose, we computed *salience scores* based on the number of news events and the two external measures, by taking the log and min-max scaling the values to the interval $[0, 1]$. Thus, for each measure, countries were mapped to a continuous scale ranging from 0 (least salient) to 1 (most salient).

If newsworthiness was proportionally distributed across countries, those that score higher (lower) in landslide occurrence should also score higher (lower) in observed news events. Deviations from that proportional relation suggest that a country has been receiving more or less media attention than expected. To operationalise this rationale, we computed the residuals of a linear regression as a proxy for salience divergence, using news salience as the dependent variable and external salience as the independent variable:

$$divergence = salience_{news} - (\hat{\beta}_0 + \hat{\beta}_1 salience_{ext})$$

where $\hat{\beta}_{\{0,1\}}$ are estimated parameters. Therefore, the difference between the observed news salience score and the prediction by the linear regression was interpreted as a deviation from what could be expected based on external measures.

Divergence close to 1 (-1) means that a country scores higher (lower) in news salience than in EM-DAT or WB-GLHM salience. Values near 0 indicate that news salience is close to what is expected. A similarity threshold was set using the magnitude of the most extreme observations to select a symmetric portion of 25% of the interval around 0, considered the fitting (i.e. non-divergent) salience range. Based on that, we grouped countries into three categories: those with news salience higher (overreported), similar or lower (underreported) compared to the external salience.

country	news events	docs	active days
Italy	314	4,991	631
Switzerland	306	3,640	638
Austria	237	3,617	521
China	217	4,438	580
USA	191	2,172	356
India	173	2,853	382
Philippines	153	4,244	411
Nepal	142	1,662	312
Brazil	138	3,056	356
Indonesia	136	3,059	291
France	120	953	175
Spain	116	974	173
Japan	114	3,045	275
Peru	103	733	166
Colombia	100	1,161	150
Pakistan	89	890	172
Norway	80	962	118
Czechia	74	158	96
UK/N. Ireland	67	443	103
Mexico	61	1,170	138

Table 1: The 20 countries with the most identified news events and the corresponding number of documents and active days in the period. Full table in the Appendix.



Figure 5: Number of documents for each news outlet (>500) and distribution of data instances by continent.

5 Results and Analysis

First, we present the estimated properties of German news about international landslides. Then, we compare our measures to external sources.

5.1 Overview: Landslides in German News

The resulting dataset contains 54,806 documents published by 239 news outlets with 25,294 text types. Figure 4 is an overview of the complete count time series depicting their temporal distribution by date. 5,098 active days occur in the 9,132-day period. Almost 45% of the days have none and 16.7% have only one document. 52 days are peaks with at least 100 documents, with maximum 217.

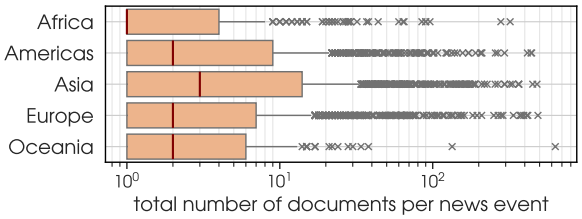
In total, there are 59,997 data instances covering 152 countries (after discarding 209 occurrences of 24 spurious ISO codes generated by the LLM) whose spatial distribution is shown in Figure 3. 92.8% of the documents are about one country and 5.4% about two. Documents concentrate in a few countries: the top ten (shown in Table 1) account for 58.5% of the total number of instances. Only 28 countries have 500 or more instances, covering 83.1% of the total; 92 countries have none and 45 have only up to 10 instances. Country-specific series are extremely sparse: the first, second and third quartiles for the number of active days are 3, 14 and 53.25, respectively. The most frequent country has 638 active days (still <7% of all days).

Figure 5 shows the number of documents for the most frequent outlets and their distribution per continent. Documents about Africa and Oceania occur less often across all news outlets. Asia is the most represented continent, followed by Europe and the Americas, with a few exceptions.

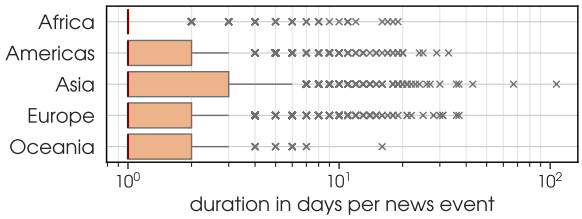
5.2 Spatio-Temporal Distribution of News Events

The segmentation of the geolocated news reporting on landslides led to the identification of 4,567 news events. Countries with the largest number of news events are listed in Table 1. The rankings for number of events, documents and active days are not the same for all countries, making these three measures complementary. Some countries have more documents forming fewer news events (e.g. China) or more news events with fewer documents (e.g. USA); similarly, some countries have more documents spread across fewer days (e.g. Indonesia) and vice versa (e.g. France).

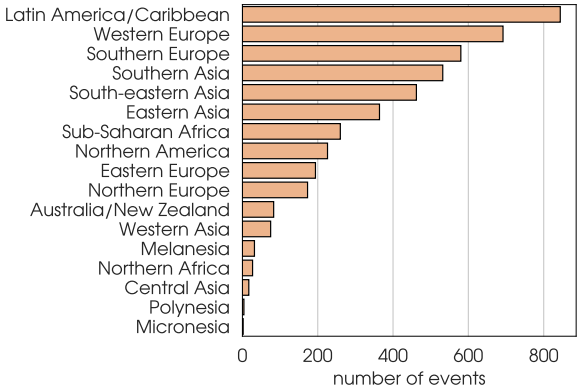
The duration and volume measures help capture such variations. The log-plots in Figures 6a and 6b show that the median number of days per news event is 1 for all continents and the median number of documents is between 1 and 3. Most news events are brief (1-3 days) with no more than 14 documents, but several outliers capture more media attention. Figures 6c and 6d show the spatial and temporal distribution of identified news events. Despite some yearly variation, there is no clear bias towards recent years as in the daily counts.



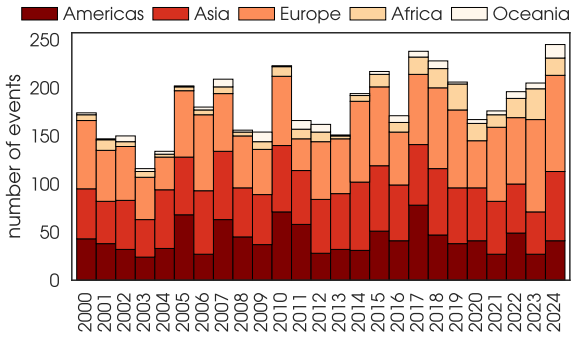
(a) Log-plot of the distribution of volume by continent.



(b) Log-plot of the distribution of duration by continent.



(c) Distribution of identified news events by subregion.



(d) Distribution of identified news events by year and continent.

Figure 6: Overview of identified news events.

5.3 Comparison to External Sources

We investigate the extent to which the non-uniform distribution of news events over countries and the resulting salience scores are due to media bias or to the real occurrence of more landslides.

Distribution of total news events In Table 2, the percentage of news events assigned to each country category is compared with the corresponding percentages in EM-DAT and WB-GLHM. In terms of hazard risk, our observations are as expected: most news events refer to countries classified as high risk, and the values are comparable to the external sources. However, the proportion of news events for the Global North is higher compared to EM-DAT and WB-GLHM, which concentrate in the Global South. This relates to continents: news events in Europe are much more frequent than in the external sources, whereas Asia is less frequent. Table 7 in the Appendix shows the values by subregion, particularly indicating the overreporting of Southern and Western Europe and underreporting of South-Eastern Asia. There is also a contrast in relation to income: news events concentrate more on high income countries. This difference comes at the cost of fewer observations than expected in the upper and lower middle income, but not so much in low income countries.

		news		% EM-DAT	% WB-GLHM
		n	%		
risk	high	4,224	92.49	93.61	97.86
	medium	217	4.75	3.29	1.14
	unknown	1	0.02	0.00	0.08
	low	89	1.95	2.95	0.75
	very low	36	0.79	0.15	0.17
develop.	global N.	2,084	45.63	13.88	24.41
	global S.	2,483	54.37	86.12	75.59
continent	Africa	287	6.28	9.39	4.16
	Americas	1,070	23.43	27.14	25.73
	Asia	1,450	31.75	53.12	57.54
	Europe	1,639	35.89	7.55	7.96
	Oceania	121	2.65	2.81	4.60
income	high	2,183	47.80	16.98	28.02
	upper mid	1,129	24.72	39.86	36.91
	unknown	48	1.05	1.31	1.13
	lower mid	1,002	21.94	33.62	30.08
	low	205	4.49	8.22	3.86

Table 2: Number and percentage of identified news events compared with external measures.

Temporal aspect We examine the distribution of news events by subregion for each year and compare the observed proportions with their external counterparts. The distribution of differences is displayed in Figure 7 for WB-GLHM (the equivalent plot for EM-DAT is in the Appendix). The yearly proportions generally deviate by only up

to 5% in relation to the external measures. Noticeable deviations occur for South-eastern Asia and Northern America, and partially Eastern Asia, with lower yearly proportions, and for Southern and Western Europe, as well as partially Latin America/Caribbean, with larger yearly proportions. The comparison to EM-DAT is mostly similar, except for Northern America, which is within the 5% range, and for opposite results for Latin America/Caribbean. Detailed plots by year are in Figure 13 (Appendix).

Salience and divergence scores The divergence scores for all countries are displayed in Figures 8a and 8c, contrasting negative and positive divergence. The scatter plots in Figures 8b and 8d help visualise how salience scores are distributed around the fitted line. For EM-DAT, the line coefficient and intercept are 0.91 and 0.06, respectively; for WB-GLHM, the estimates are 0.79 and 0.01, respectively. The grey area represents the computed similarity range. Data points above (below) it are countries considered to be overreported (underreported) in German news. Exact values by country are presented in the Table 8 (Appendix).

The percentage of observations in each divergence group is displayed in Table 3 by development level. Most countries fall within similar ranges on both external measures, corroborating the overall validity of the dataset. Still, some countries both in the Global North and South are more/less salient in the news than expected. A considerable portion of countries in the Global North are more salient in the news than expected and only two are underreported (Albania and North Macedonia) based on both external measures, whereas the portion of countries in the Global South that are less salient than expected is higher for WB-GLHM than EM-DAT.

The most extreme underreporting (negative divergence) with respect to EM-DAT occurs for Tajikistan, North Macedonia, Côte d'Ivoire, Trinidad and Tobago, and Congo; the overreporting (positive divergence) extremes are Spain, Austria, Switzerland, Norway and UK/Northern Ireland. For WB-GLHM, the negative extremes include Mongolia, Liberia, Gabon, Uzbekistan and Azerbaijan; the positive extremes are Italy, Sri Lanka, Austria, Czechia and Switzerland. Only two countries fall into opposite extremes between EM-DAT and WB-GLHM: Cambodia (higher for EM-DAT but lower for WB-GLHM) and Rwanda (lower for EM-DAT but higher for WB-GLHM).

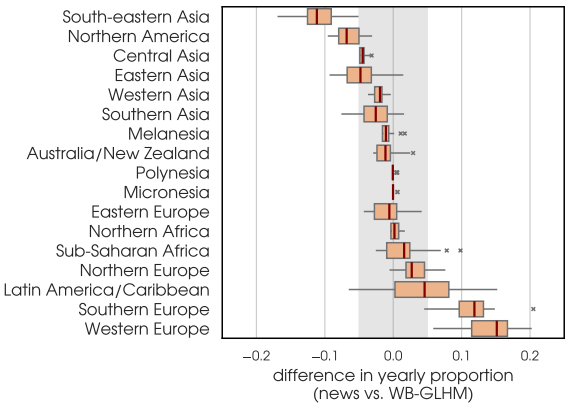


Figure 7: Box-plot of the differences in yearly proportion between news events and WB-GLHM.

		lower in news	similar scores	higher in news
news vs. EM-DAT	Global North	3.12	53.12	43.75
	Global South	11.67	85.00	3.33
news vs. WB-GLHM	Global North	3.12	65.62	31.25
	Global South	21.67	65.56	12.78

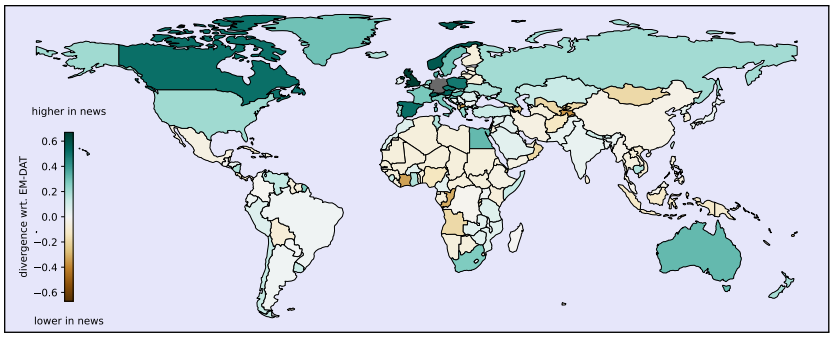
Table 3: % of under/overreported countries.

6 Discussion

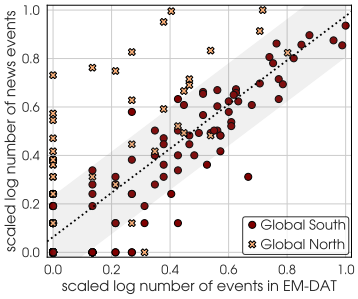
According to our estimations, German newspapers discussed landslides abroad more than 4.5k times in 2000-2024. Although a few countries were more salient than others in the absolute number of documents and news events, part of this imbalance followed expected patterns of susceptibility to landslides, as evidenced by historical data of observed events (EM-DAT) and a data-driven predictive model (WB-GLHM). We have validated the resulting dataset in terms of overall distribution by a country-level indicator of hazard risk and by salience scores, with some interesting deviations.

Some countries/subregions turned out to be disproportionally salient in German news: there were more news events about countries in the Global North, mostly in some regions of Europe and/or of those of high income, than expected if newsworthiness was proportionally distributed based on countries' susceptibility to landslides. Conversely, a few countries have received less attention than expected, including parts of Asia and countries with middle income levels.

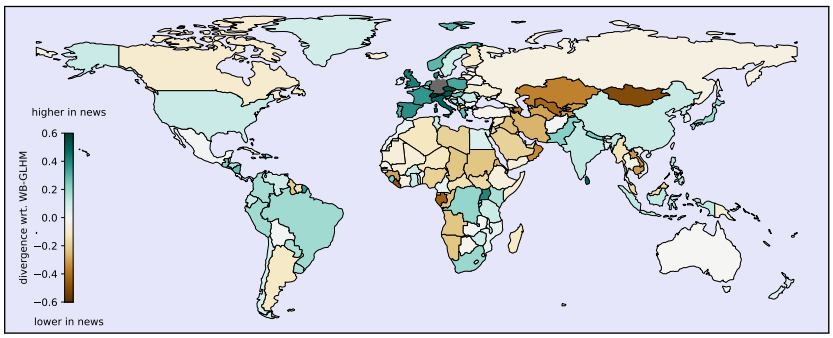
These asymmetries shed light on systematic biases in disaster reporting that demand further investigation. The particular media bias explored here suggests that German-speaking audiences receive



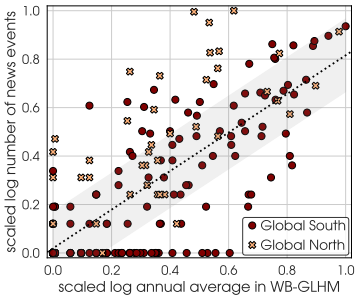
(a) Divergence scores: EM-DAT vs. news events



(b) Salience: EM-DAT vs. news.



(c) Divergence scores: WB-GLHM vs. news events.



(d) Salience: WB-GLHM vs. news.

Figure 8: Divergence scores and linear regression with the defined similarity boundary (in grey) for salience scores.

a partially distorted picture of global landslide risk. A skewed coverage has implications for climate awareness and preparedness: it may limit public understanding of the true scale and geographic distribution of disasters, potentially reducing political will and public support for international climate adaptation financing and risk reduction initiatives.

The different results in some of the comparisons to EM-DAT and WB-GLHM ought to derive from their distinct nature: EM-DAT only considers observed events that made their way to a few curated sources (such as selected press agencies, governmental agencies and insurance companies) and surpassed a given impact level. Due to geographical and cultural proximity, German newspapers possibly often report on local, less severe events in Europe that do not meet such criteria (e.g. roads blocked by landslides), which could explain the higher proportion of news events in Europe. The WB-GLHM model, on the other hand, is based on real-world environmental observations. It should account for all types of events, despite their severity; still, it is merely *predictive* and may misestimate actual events in each year. Canada, for example, is the 12th most salient country in WB-GLHM but had only 1 entry in EM-DAT, leading to opposite signs in its divergence scores.

7 Conclusions and Future Work

We analysed a large-scale dataset of German news articles about landslides worldwide which can enrich incomplete disaster databases and enable large-N studies of inequalities in media attention. Our results revealed preliminary insights regarding imbalance in coverage relative to a global landslide inventory (EM-DAT) and a predictive landslide frequency model (WB-GLHM). Overall, news in German about landslides tend to favour some countries and regions while overlooking others. It stands as a foundational layer towards understanding how social inequalities are reflected and reinforced by disaster coverage in Global North's media systems.

This work paves the way for a variety of relevant research questions. Possible next steps are modelling media coverage based on countries' socio-economic, cultural, linguistic and geographical proximity to Germany; assessing the temporal dynamics of media attention cycles; distinguishing different types of news events (i.e. reporting about past, ongoing or future/hypothetical events, as in Huang et al., 2016); employing alignment methods and information extraction techniques to extract event metadata; expanding the analysis to other hazards such as wildfires and floods; and examining the language use for reporting on landslides.

Limitations

The LLM performance did not generalise to other types of hazards like cold waves and droughts. The prompts should thus not be directly used for other datasets without careful evaluation. Although the relevance classification and geolocation performance were high, some errors still occur, making unrelated and/or wrongly geolocated documents form spurious news events. False positives can be a problem especially for countries that are rarely observed and have few news events. Moreover, LLMs can introduce biases in geocoding, which, in turn, influence results and conclusions. A stratified analysis of the LLM's geolocation codes has not been conducted yet.

We did not try to distinguish texts that only *mention* a landlside among other topics from texts whose main purpose is to report on a landslide, such as the ones in Figure 1. Depending on the purpose of the study using this dataset, this distinction may be required.

We inherited some known and unknown biases from the `wiso-net` database. We assumed it is representative of German newspapers, but their news extraction pipeline may not have included all relevant news documents and/or news outlets. Some documents are actually a concatenation of many pieces of news, without clear distinctive boundaries, which may negatively impact classification and geolocation.

The minimum number of inactive days between different news events is a hyperparameter. Other values apart from 5 can be tested. Lower values lead to more news events, possibly splitting the same topic into more than one event, whereas higher values may cause different events to be merged into the same news event.

We did not treat the temporal bias toward recent years in the number of documents here because it is not so prominent after event segmentation and because we mainly focused on the total number of events in the period. Still, depending on the type of analysis, normalisation methods for the number of documents may be necessary.

The spatial analysis was performed only at the country level; we neither located events regionally nor normalised counts by area. Refining this aspect can be particularly relevant for large countries like Brazil, US and China.

The validation with external data sources was performed using the EM-DAT inventory, a well-established source in the field (despite its known limitations and biases) that covers the full period of our observations and can easily be extended to other hazards in our pipeline. Future work can integrate and compare results using the Unified Global Landslide Database by Gómez et al. (2023) (although with observations only up to 2020) and the the Global Fatal Landslide Database by Froude and Petley (2018) (only from 2004 to 2016).

Data and Code

We do not have permission to share the text data; it must be accessed via the wiso-net database at `https://www.wiso-net.de/`. Details about the specific used instances are available upon request. The code used to produce the results presented in this paper is available at `https://codeberg.org/briemadu/landslides-german-news`.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions.

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.

Shayan Alipour, Niccolò Di Marco, Michele Avalle, Gabriele Etta, Matteo Cinelli, and Walter Quattrociocchi. 2024. The drivers of global news spreading patterns. *Scientific Reports*, 14(1).

Aydoğan Avcıoğlu, Ogün Demir, and Tolga Görüm. 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(7):2421–2435.

Alessandro Battistini, Ascanio Rosi, Samuele Segoni, Daniela Lagomarsino, Filippo Catani, and Nicola Casagli. 2017. Validation of landslide hazard models using a semantic engine on online news. *Applied Geography*, 82:59–65.

Tal Bauman, Yuval Fisher, Fadi Kizel, Itzhak Omer, and Sagi Dalyot. 2025. Geographic information retrieval: a systematic review of datasets and their impact on textual geolocation. *Geo-spatial Information Science*, page 1–18.

José Eduardo Bonini, Tiago Damas Martins, Marina Tamaki de Oliveira Sugiyama, and Bianca Carvalho

Vieira. 2025. Landslide inventory of the 2023 serra do mar disaster (brazil). *Discover Geoscience*, 3(1).

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.

Richard A. Davis, Konstantinos Fokianos, Scott H. Holan, Harry Joe, James Livsey, Robert Lund, Vladas Pipiras, and Nalini Ravishanker. 2021. Count time series: A methodological review. *Journal of the American Statistical Association*, 116(535):1533–1547.

Damien Delforge, Valentin Wathelet, 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.

María José Domínguez Cuesta, Montserrat Jiménez Sánchez, and Augusto Rodríguez García. 1999. Press archives as temporal records of landslides in the north of spain: relationships between rainfall and instability slope events. *Geomorphology*, 30(1):125–132.

Harideva Marturano Egas, Rodrigo Augusto Stabile, Marcio Roberto Magalhães de Andrade, Gean Paulo Michel, João Paulo Carvalho de Araújo, Rossano Dalla Lana Michel, Tatiana Sussel Gonçalves Mendes, Tulius Dias Nery, Danielle Silva de Paula, and Elisabete Weber Reckziegel. 2024. Comprehensive inventory and initial assessment of landslides triggered by autumn 2024 rainfall in rio grande do sul, brazil. *Landslides*, 22(2):579–589.

Robert Emberson, Dalia Kirschbaum, and Thomas Stanley. 2020. New global characterisation of landslide exposure. *Natural Hazards and Earth System Sciences*, 20(12):3413–3424.

Rachele Franceschini, Ascanio Rosi, Filippo Catani, and Nicola Casagli. 2022. Exploring a landslide inventory created by automated web data mining: the case of italy. *Landslides*, 19(4):841–853.

Melanie J. Froude and David N. Petley. 2018. Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18(8):2161–2181.

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.

Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news. *Journal of Peace Research*, 2(1):64–90.

Stefano Luigi Gariano and Fausto Guzzetti. 2022. *Mass-Movements and Climate Change*, page 546–558. Elsevier.

Derly Gómez, Edwin F. García, and Edier Aristizábal. 2023. Spatial and temporal landslide distributions using global and open landslide databases. *Natural Hazards*, 117(1):25–55.

Lynn M Highland and Peter Bobrowsky. 2008. *The landslide handbook - A guide to understanding landslides*. Circular 1325. U.S. Geological Survey.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The EventStatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44–54, Austin, Texas. ACL.

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.

Stijn Joye. 2010. On the media construction of international disasters. *Tijdschrift voor Communicatiewetenschap*, 38(2):139–155.

Dalia Bach Kirschbaum, Robert Adler, Yang Hong, Stephanie Hill, and Arthur Lerner-Lam. 2009. A global landslide catalog for hazard applications: method, results, and limitations. *Natural Hazards*, 52(3):561–575.

Oddvar Kjekstad and Lynn Highland. 2009. *Economic and Social Impacts of Landslides*, page 573–587. Springer Berlin Heidelberg.

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.

Thomas M. Kreuzer and Bodo Damm. 2020. Automated digital data acquisition for landslide inventories. *Landslides*, 17(9):2205–2215.

Lukas Kriesch and Sebastian Losacker. 2025. A geolocated dataset of German news articles. *Scientific Data*, 12(1):1128.

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. ACL.

Brielen Madureira, Mariana Madruga de Brito, and Andreas Niekler. 2026a. Retrieving floods without floodlights: Topic models as binary classifiers for extreme climate events in German news. In *Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing*, pages 115–134, Palma de Mallorca, Spain. European Language Resources Association.

Brielen Madureira, Andreas Niekler, Marc Keuschnigg, and Mariana Madruga de Brito. 2026b. The newsworthiness of brazilian distress: A peak analysis on time series of international media attention to disasters in brazil. *Preprint*, arXiv:2605.04552.

Tapas Ranjan Martha, Priyom Roy, Nirmala Jain, Kirti Khanna, K. Mrinalni, K. Vinod Kumar, and P. V. N. Rao. 2021. Geospatial landslide inventory of india—an insight into occurrence and exposure on a national scale. *Landslides*, 18(6):2125–2141.

Chris Massey, Kerry Leith, Tom R Robinson, Biljana Lukovic, Sam McColl, Trevor Carey-Smith, Brenda Rosser, Liam Wotherspoon, Hugh Smith, Harley Betts, Rob Buxton, and Janine Bidmead. 2025. What controlled the occurrence of more than 116, 000 human-mapped landslides triggered by cyclone gabrielle, new zealand? *Landslides*, 22(12):3953–3972.

Benjamin B. Mirus, Eric S. Jones, Rex L. Baum, Jonathan W. Godt, Stephen Slaughter, Matthew M. Crawford, Jeremy Lancaster, Thomas Stanley, Dalia B. Kirschbaum, William J. Burns, Robert G. Schmitt, Kassandra O. Lindsey, and Kevin M. McCoy. 2020. Landslides across the usa: occurrence, susceptibility, and data limitations. *Landslides*, 17(10):2271–2285.

Kieu Anh Nguyen, Yi-Jia Jiang, Chiao-Shin Huang, Meng-Hsun Kuo, and Walter Chen. 2024. Leveraging internet news-based data for rockfall hazard susceptibility assessment on highways. *Future Internet*, 16(8):299.

Priscilla Niyokwiringirwa, Luigi Lombardo, Olivier Dewitte, Axel A. J. Deijns, Nan Wang, Cees J. Van Westen, and Hakan Tanyas. 2024. Event-based rainfall-induced landslide inventories and rainfall thresholds for malawi. *Landslides*, 21(6):1403–1424.

Einar Östgaard. 1965. Factors influencing the flow of news. *Journal of Peace Research*, 2(1):39–63.

David Petley. 2012. Global patterns of loss of life from landslides. *Geology*, 40(10):927–930.

Karl Erik Rosengren. 1970. International news: Intra and extra media data. *Acta Sociologica*, 13(2):96–109.

Denise J. Roth, Nan Bai, Robbert Biesbroek, Art Dewulf, Sanne Kruikemeier, Daan de Leur, Mariken A. C. G. van der Velden, Erik de Vries, and Rens Vliegenthart. 2025. Unveiling climate adaptation: A comprehensive analysis of agenda-setting dynamics in the united kingdom and the netherlands. *Science Communication*.

Jan Sodoge, 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.

Faith E. Taylor, Bruce D. Malamud, Katy Freeborough, and David Demeritt. 2015. Enriching great britain's national landslide database by searching newspaper archives. *Geomorphology*, 249:52–68.

Bram Valkenborg, Olivier Dewitte, and Benoît Smets. 2026. Unravelling information on impactful geohydrological hazard events with hazminer, a multilingual text mining method developed through a global scale coverage application. *EGUsphere*, 2026:1–45.

Miet Van Den Eeckhaut, Javier Hervás, and Luca Montanarella. 2013. *Landslide Databases in Europe: Analysis and Recommendations for Interoperability and Harmonisation*, page 35–42. Springer Berlin Heidelberg.

Chenchen Xie, Yuandong Huang, Chong Xu, Kebin Dai, and Xiwei Xu. 2026. Over 100, 000 landslides triggered by typhoon-induced rainfall in north china in july 2023. *Landslides*, 23(5):1389–1408.

M. M. Yagoub, Aishah A. Alsereidi, Elfadil A. Mohamed, Punitha Periyasamy, Reem Alameri, Salama Aldarmaki, and Yaqein Alhashmi. 2020. Newspapers as a validation proxy for gis modeling in fujairah, united arab emirates: identifying flood-prone areas. *Natural Hazards*, 104(1):111–141.

Yan Yan and Kim Bissell. 2015. The sky is falling: Predictors of news coverage of natural disasters worldwide. *Communication Research*, 45(6):862–886.

A Appendix

A.1 Processing Details

The list of German keywords used to query news documents in shown in Table 4.

The LLM was deployed via LM Studio.[3] The system prompt for classification was:

```
You are an expert in reading news articles carefully and
    classifying them very precisely based on their
    references to $hazard events.
```

The main prompt for classification was:

```
The German news article below must be classified with one
    of these labels:

– Label 1: At least one sentence in the text refers to a $
    hazard event.
– Label 0: The text does not refer to any $hazard event.

Some keywords for $hazard in German are: $keywords.

Only concrete, specific and natural $hazard events that
    happenned or are happenning in the real world are
    relevant. Articles that mention only the possibility or
    risk of such an event and general discussions about $
    hazardplural with no reference to an specific event
    must be ignored.

Decide which the correct label is and begin your answer
    with 0 or 1 accordingly. Do not begin your answer
    with any other word or symbol.

Here is the news article:

$text
```

The system prompt for geolocation was:

```
You are a very precise information extractor that identifies
    the country where $hazard occurred as reported in
    news articles.
You always output the results in JSON format and identify
    affected countries as ISO 3166−1 alpha−3 codes.
Example: {"country_codes": "DEU;BRA"}
```

The main prompt for geolocation was:

```
Analyze the news article below and identify the mentioned
    country or countries where $hazard occurred.
Follow these constraints:
1. Only past or ongoing $hazard are relevant. Ignore
    possible, future, hypothetical or fictional $hazard and
    all other types of events.
2. Only include countries that are stated in this news article.
    Do not use outside knowledge.
3. Return the result strictly in JSON format and convert
    country names to ISO 3166−1 alpha−3 codes.
4. If no countries are mentioned as being affected by $
    hazard, return N/A for the country_codes value.
5. If no past or ongoing $hazard are described, return N/A
    for the country_codes value.
6. Do not include any country for which only other events
    apart from $hazard are discussed in the text.

Here is the news article:
$text
```

[3] https://lmstudio.ai/

German	English
Erdrutsch	Landslide
Felssturz, Felsstürz	Rockfall
Schlammlawine	Mudslide
Massenbewegung	Mass movement
Hangrutsch	Slope slide
Hangbewegung	Slope movement
Rutschung	Slip
Bodenrutsch	Ground slide
Hangabrutschung	Hillside slumping
Murgang	Debris flow
Gerölllawine	Debris avalanche
Rutschhang, Rutschhäng	Sliding slope
Rutschgefahr	Slip hazard
Felslawine	Rock avalanche
Mure	Mudflow

Table 4: German keywords used to query the news database and their (approximate) translations.

Countries Data instances about Germany were ignored, as we are only interested in events in other countries. Taiwan was added to the reference list of countries as it appears in the predictions and in country indicators. Antarctica, Åland Islands, United States Minor Outlying Islands and Niue were excluded due to lack of data. The countries' polygon coordinates were retrieved from the Natural Earth Data project. The download links for all country indicators are listed in Table 5.

For computing the distributions in Figure 5, documents about more than one continent were counted as separate instances.

A.2 Additional Results

Figures 9 and 10 show the number of identified news events and active days per country (complementing Figure 3). Figure 11 shows the almost linear relation between document and news events; generally, more documents result in more events, but a few exceptions occur, e.g. countries with just one event that encompass a larger number of documents. Figure 12 shows EM-DAT yearly distributions as a counterpart of Figure 7. The subplots in Figure 13 represent the evolution of the proportion of documents per subregion over the time period.

Table 7 extends the information in Table 2 with subregions. Finally, Table 8 lists the computed divergence scores for all countries in the sample.

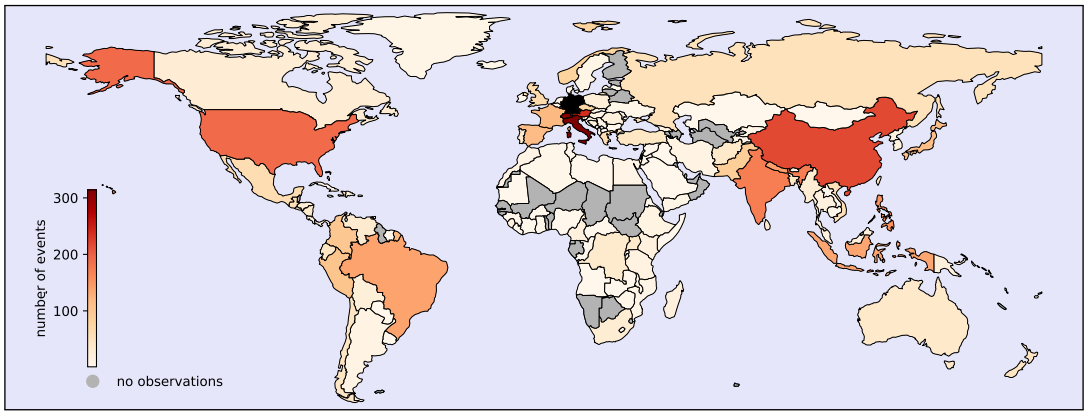


Figure 9: Total number of news events about landslides in each country. Germany (in black) was not analysed.

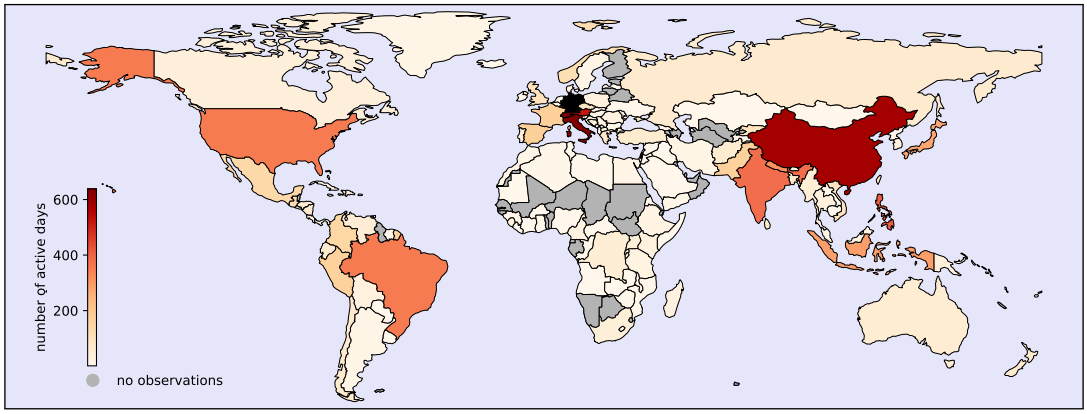


Figure 10: Total number of active days for each country. Germany (in black) was not analysed.

UNSD countries	https://unstats.un.org/unsd/methodology/m49/overview/
EM-DAT	https://www.emdat.be/
WB-GLHM	https://datacatalog.worldbank.org/search/dataset/0037584/global-landslide-hazard-map
TH-Risk	https://datacatalog.worldbank.org/search/dataset/0060151/thinkhazard-hazard-ranking
Development status	https://unctadstat.unctad.org/EN/Classifications.html
Income group	https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
Polygons	http//www.naturalearthdata.com/download/110m/cultural/ne_110m_admin_0_countries.zip

Table 5: Download links for the external data sources.

measure	**description**
n at peak	daily article count at the event's peak date
total volume	total number of articles during the news event
duration	total number of days in the news event
days since last	number of days since the last news event
days to peak	number of days from the first article's date until the peak date
days to fade	number of days from the peak date to the last article's date
n text types	number of unique text types published during the news event
n outlets	number of media outlets reporting during the news event

Table 6: List of all measures computed for the identified news events, based on (Madureira et al., 2026b).

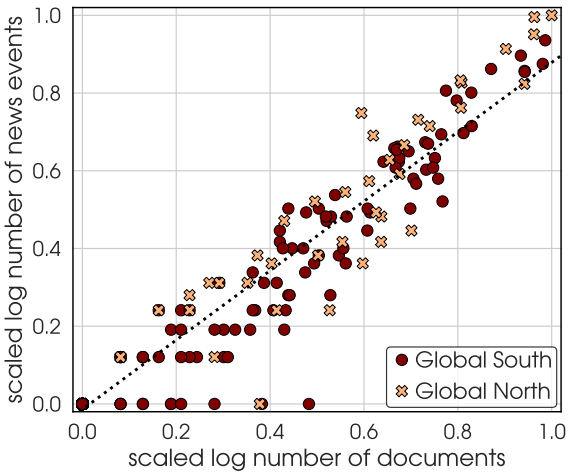


Figure 11: Relation between number of documents and number of news events.

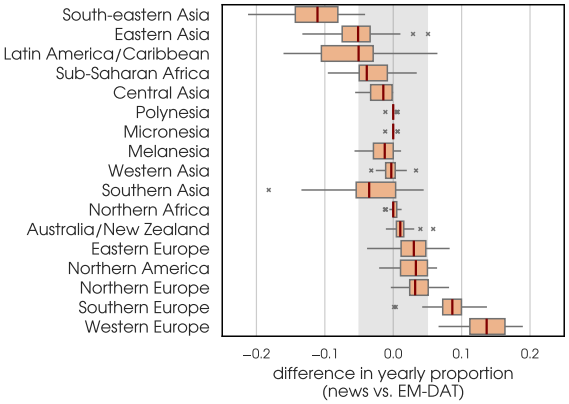


Figure 12: Box-plot of the differences in yearly proportion of news events and EM-DAT.

subregion	news n	news %	% EM-DAT	% WB-GLHM
Northern Africa	27	0.59	0.48	0.34
Sub-Saharan Africa	260	5.69	8.90	3.83
Latin America/Caribbean	844	18.48	25.25	14.31
Northern America	226	4.95	1.89	11.42
Central Asia	17	0.37	2.37	4.88
Eastern Asia	364	7.97	12.58	13.46
South-eastern Asia	462	10.12	21.09	21.30
Southern Asia	532	11.65	15.05	14.20
Western Asia	75	1.64	2.03	3.69
Eastern Europe	194	4.25	1.31	5.18
Northern Europe	173	3.79	0.15	0.54
Southern Europe	580	12.70	4.45	1.46
Western Europe	692	15.15	1.64	0.78
Australia/New Zealand	83	1.82	0.68	2.93
Melanesia	32	0.70	2.03	1.58
Micronesia	2	0.04	0.05	0.03
Polynesia	4	0.09	0.05	0.06

Table 7: Comparison of the number and percentage of identified news events to external measures (subregions).



Figure 13: Proportion of (news) events by subregion for each year.

country	documents	active days	news events	divergence EM-DAT	divergence WB-GLHM	binned score EM-DAT	binned score WB-GLHM
ABW	1	1	1	-0.063	-0.019	similar scores	similar scores
AFG	600	61	38	-0.134	0.008	similar scores	similar scores
AGO	1	1	1	-0.186	-0.229	lower in news	lower in news
AIA	0	0	0	-0.063	-0.020	similar scores	similar scores
ALB	2	2	2	-0.188	-0.235	lower in news	lower in news
AND	0	0	0	-0.063	-0.020	similar scores	similar scores
ARE	0	0	0	-0.063	-0.154	similar scores	lower in news
ARG	55	14	10	0.020	-0.115	similar scores	similar scores
ARM	0	0	0	-0.063	-0.287	similar scores	lower in news
ASM	0	0	0	-0.063	-0.020	similar scores	similar scores
ATF	0	0	0	-0.063	-0.020	similar scores	similar scores
ATG	0	0	0	-0.063	-0.020	similar scores	similar scores
AUS	264	56	37	0.319	-0.003	higher in news	similar scores
AUT	3,617	521	237	0.543	0.507	higher in news	higher in news
AZE	0	0	0	-0.258	-0.383	lower in news	lower in news
BDI	22	10	7	-0.114	0.141	similar scores	similar scores
BEL	20	9	6	0.126	0.216	similar scores	higher in news
BEN	0	0	0	-0.063	-0.025	similar scores	similar scores
BES	0	0	0	-0.063	-0.020	similar scores	similar scores
BFA	8	2	2	0.057	0.101	similar scores	similar scores
BGD	372	63	42	0.024	0.054	similar scores	similar scores
BGR	7	6	5	0.022	0.004	similar scores	similar scores
BHR	0	0	0	-0.063	-0.020	similar scores	similar scores
BHS	2	2	1	-0.063	-0.025	similar scores	similar scores
BIH	206	40	17	0.022	0.155	similar scores	higher in news
BLM	0	0	0	-0.063	-0.020	similar scores	similar scores
BLR	0	0	0	-0.063	-0.020	similar scores	similar scores
BLZ	0	0	0	-0.063	-0.166	similar scores	lower in news
BMU	0	0	0	-0.063	-0.020	similar scores	similar scores
BOL	98	35	22	-0.081	0.005	similar scores	similar scores
BRA	3,056	356	138	0.020	0.183	similar scores	higher in news
BRB	0	0	0	-0.063	-0.020	similar scores	similar scores
BRN	0	0	0	-0.063	-0.096	similar scores	similar scores
BTN	42	14	5	0.094	-0.109	similar scores	similar scores
BVT	0	0	0	-0.063	-0.020	similar scores	similar scores
BWA	0	0	0	-0.063	-0.020	similar scores	similar scores
CAF	5	2	1	-0.063	-0.137	similar scores	similar scores
CAN	182	43	27	0.510	-0.091	higher in news	similar scores
CCK	0	0	0	-0.063	-0.020	similar scores	similar scores
CHE	3,640	638	306	0.564	0.592	higher in news	higher in news
CHL	306	66	45	0.131	0.079	similar scores	similar scores
CHN	4,438	580	217	-0.039	0.120	similar scores	similar scores
CIV	14	3	2	-0.332	-0.045	lower in news	similar scores
CMR	42	26	18	0.050	0.029	similar scores	similar scores
COD	312	51	36	-0.017	0.204	similar scores	higher in news
COG	2	1	1	-0.309	-0.241	lower in news	lower in news
COK	0	0	0	-0.063	-0.020	similar scores	similar scores
COL	1,161	150	100	-0.012	0.177	similar scores	higher in news
COM	1	1	1	-0.186	-0.307	lower in news	lower in news
CPV	0	0	0	-0.186	-0.020	lower in news	similar scores
CRI	91	22	16	-0.103	-0.022	similar scores	similar scores
CUB	27	11	6	-0.096	-0.010	similar scores	similar scores
CUW	0	0	0	-0.063	-0.020	similar scores	similar scores
CXR	0	0	0	-0.063	-0.020	similar scores	similar scores
CYM	3	2	2	0.057	0.101	similar scores	similar scores
CYP	0	0	0	-0.063	-0.096	similar scores	similar scores
CZE	158	96	74	0.491	0.520	higher in news	higher in news
DJI	0	0	0	-0.186	-0.198	lower in news	lower in news
DMA	57	11	7	0.152	0.319	similar scores	higher in news
DNK	226	22	11	0.354	0.398	higher in news	higher in news
DOM	185	28	17	-0.062	0.209	similar scores	higher in news
DZA	40	8	4	-0.068	-0.125	similar scores	similar scores
ECU	287	66	44	0.025	0.116	similar scores	similar scores
EGY	105	17	9	0.319	0.069	higher in news	similar scores
ERI	1	1	1	-0.063	-0.267	similar scores	lower in news
ESH	0	0	0	-0.063	-0.020	similar scores	similar scores
ESP	974	173	116	0.518	0.380	higher in news	higher in news
EST	0	0	0	-0.063	-0.020	similar scores	similar scores
ETH	176	21	13	-0.042	-0.058	similar scores	similar scores
FIN	0	0	0	-0.063	-0.096	similar scores	similar scores
FJI	22	4	4	-0.191	-0.107	lower in news	similar scores
FLK	0	0	0	-0.063	-0.020	similar scores	similar scores
FRA	953	175	120	0.278	0.361	higher in news	higher in news
FRO	0	0	0	-0.063	-0.020	similar scores	similar scores
FSM	2	2	1	-0.063	-0.019	similar scores	similar scores
GAB	0	0	0	-0.063	-0.440	similar scores	lower in news
GBR	443	103	67	0.668	0.421	higher in news	higher in news
GEO	114	15	10	-0.103	-0.054	similar scores	similar scores
GGY	0	0	0	-0.063	-0.020	similar scores	similar scores
GHA	6	6	4	0.178	0.104	higher in news	similar scores
GIB	0	0	0	-0.063	-0.020	similar scores	similar scores
GIN	11	1	1	-0.063	-0.298	similar scores	lower in news
GLP	0	0	0	-0.063	-0.025	similar scores	similar scores

GMB	1	1	1	-0.063	-0.019	similar scores	similar scores
GNB	0	0	0	-0.063	-0.020	similar scores	similar scores
GNQ	0	0	0	-0.063	-0.257	similar scores	lower in news
GRC	546	87	61	0.227	0.278	higher in news	higher in news
GRD	0	0	0	-0.186	-0.020	lower in news	similar scores
GRL	31	13	8	0.298	0.091	higher in news	similar scores
GTM	1,008	122	55	-0.011	0.241	similar scores	higher in news
GUF	0	0	0	-0.063	-0.096	similar scores	similar scores
GUM	0	0	0	-0.063	-0.025	similar scores	similar scores
GUY	0	0	0	-0.063	-0.206	similar scores	lower in news
HKG	39	3	3	0.128	0.124	similar scores	similar scores
HMD	0	0	0	-0.063	-0.020	similar scores	similar scores
HND	427	53	26	0.036	0.156	similar scores	higher in news
HRV	24	14	9	0.319	0.078	higher in news	similar scores
HTI	503	92	48	0.040	0.379	similar scores	higher in news
HUN	72	18	9	0.319	0.264	higher in news	higher in news
IDN	3,059	291	136	-0.110	0.128	similar scores	similar scores
IMN	0	0	0	-0.063	-0.020	similar scores	similar scores
IND	2,853	382	173	0.034	0.127	similar scores	similar scores
IOT	0	0	0	-0.063	-0.020	similar scores	similar scores
IRL	11	3	2	0.057	-0.016	similar scores	similar scores
IRN	38	15	10	-0.052	-0.263	similar scores	lower in news
IRQ	4	2	2	0.057	-0.248	similar scores	lower in news
ISL	7	5	4	0.178	-0.069	higher in news	similar scores
ISR	12	6	6	0.248	0.193	higher in news	higher in news
ITA	4,991	631	314	0.283	0.488	higher in news	higher in news
JAM	45	11	10	0.020	0.165	similar scores	higher in news
JEY	0	0	0	-0.063	-0.020	similar scores	similar scores
JOR	1	1	1	-0.063	-0.166	similar scores	lower in news
JPN	3,045	275	114	0.029	0.188	similar scores	higher in news
KAZ	21	4	3	0.128	-0.359	similar scores	lower in news
KEN	73	32	18	-0.073	0.165	similar scores	higher in news
KGZ	119	99	8	-0.181	-0.308	lower in news	lower in news
KHM	7	4	4	0.178	-0.311	higher in news	lower in news
KIR	1	1	1	-0.063	-0.019	similar scores	similar scores
KNA	0	0	0	-0.063	-0.020	similar scores	similar scores
KOR	227	33	16	-0.072	0.014	similar scores	similar scores
KWT	0	0	0	-0.063	-0.020	similar scores	similar scores
LAO	23	4	4	-0.068	-0.318	similar scores	lower in news
LBN	3	1	1	-0.063	-0.207	similar scores	lower in news
LBR	5	1	1	-0.063	-0.462	similar scores	lower in news
LBY	26	2	1	-0.063	-0.188	similar scores	lower in news
LCA	0	0	0	-0.186	-0.020	lower in news	similar scores
LIE	2	2	2	0.057	0.101	similar scores	similar scores
LKA	581	66	33	-0.075	0.490	similar scores	higher in news
LSO	2	2	2	0.057	0.054	similar scores	similar scores
LTU	0	0	0	-0.063	-0.020	similar scores	similar scores
LUX	39	16	15	0.408	0.446	higher in news	higher in news
LVA	1	1	1	-0.063	-0.019	similar scores	similar scores
MAC	0	0	0	-0.063	-0.020	similar scores	similar scores
MAF	0	0	0	-0.063	-0.020	similar scores	similar scores
MAR	73	14	9	0.073	0.038	similar scores	similar scores
MCO	0	0	0	-0.063	-0.020	similar scores	similar scores
MDA	2	2	2	0.057	0.101	similar scores	similar scores
MDG	121	27	16	-0.006	-0.105	similar scores	similar scores
MDV	0	0	0	-0.063	-0.020	similar scores	similar scores
MEX	1,170	138	61	-0.052	0.007	similar scores	similar scores
MHL	0	0	0	-0.063	-0.020	similar scores	similar scores
MKD	25	1	1	-0.348	-0.154	lower in news	lower in news
MLI	0	0	0	-0.063	-0.067	similar scores	similar scores
MLT	1	1	1	-0.063	-0.019	similar scores	similar scores
MMR	636	51	28	-0.031	-0.131	similar scores	similar scores
MNE	34	5	4	0.178	0.012	higher in news	similar scores
MNG	1	1	1	-0.186	-0.500	lower in news	lower in news
MNP	0	0	0	-0.063	-0.020	similar scores	similar scores
MOZ	34	7	6	0.054	0.032	similar scores	similar scores
MRT	1	1	1	-0.063	-0.067	similar scores	similar scores
MSR	0	0	0	-0.063	-0.020	similar scores	similar scores
MTQ	1	1	1	-0.186	-0.019	lower in news	similar scores
MUS	1	1	1	-0.063	-0.137	similar scores	similar scores
MWI	90	14	5	0.022	0.082	similar scores	similar scores
MYS	58	23	17	-0.025	-0.160	similar scores	lower in news
MYT	0	0	0	-0.063	-0.020	similar scores	similar scores
NAM	0	0	0	-0.063	-0.229	similar scores	lower in news
NCL	0	0	0	-0.063	-0.137	similar scores	similar scores
NER	0	0	0	-0.063	-0.119	similar scores	similar scores
NFK	0	0	0	-0.063	-0.020	similar scores	similar scores
NGA	2	2	2	-0.137	-0.192	similar scores	lower in news
NIC	314	64	38	0.180	0.280	higher in news	higher in news
NLD	10	6	6	0.248	0.292	higher in news	higher in news
NOR	962	118	80	0.576	0.294	higher in news	higher in news
NPL	1,662	312	142	0.102	0.246	similar scores	higher in news
NRU	0	0	0	-0.063	-0.020	similar scores	similar scores
NZL	344	66	46	0.195	0.064	higher in news	similar scores
OMN	0	0	0	-0.186	-0.344	lower in news	lower in news
PAK	890	172	89	0.032	0.227	similar scores	higher in news
PAN	83	22	16	-0.103	0.072	similar scores	similar scores

PCN	0	0	0	-0.063	-0.020	similar scores	similar scores
PER	733	166	103	0.068	0.185	similar scores	higher in news
PHL	4,244	411	153	-0.062	0.144	similar scores	similar scores
PLW	0	0	0	-0.063	-0.096	similar scores	similar scores
PNG	689	37	20	-0.098	-0.069	similar scores	similar scores
POL	118	35	23	0.482	0.324	higher in news	higher in news
PRI	2	2	2	-0.228	-0.045	lower in news	similar scores
PRK	36	18	13	0.038	0.102	similar scores	similar scores
PRT	316	39	30	0.184	0.297	higher in news	higher in news
PRY	5	3	3	0.005	0.166	similar scores	higher in news
PSE	0	0	0	-0.063	-0.096	similar scores	similar scores
PYF	0	0	0	-0.063	-0.247	similar scores	lower in news
QAT	0	0	0	-0.063	-0.020	similar scores	similar scores
REU	0	0	0	-0.186	-0.301	lower in news	lower in news
ROU	68	27	20	0.069	0.112	similar scores	similar scores
RUS	196	66	53	0.202	-0.044	higher in news	similar scores
RWA	36	14	11	-0.168	0.188	lower in news	higher in news
SAU	3	3	2	0.057	-0.189	similar scores	lower in news
SDN	0	0	0	-0.063	-0.229	similar scores	lower in news
SEN	0	0	0	-0.063	-0.020	similar scores	similar scores
SGP	0	0	0	-0.063	-0.020	similar scores	similar scores
SGS	0	0	0	-0.063	-0.025	similar scores	similar scores
SHN	0	0	0	-0.063	-0.020	similar scores	similar scores
SJM	0	0	0	-0.063	-0.025	similar scores	similar scores
SLB	16	3	3	-0.157	-0.245	similar scores	lower in news
SLE	384	34	18	0.122	0.241	similar scores	higher in news
SLV	512	75	32	0.027	0.336	similar scores	higher in news
SMR	0	0	0	-0.063	-0.020	similar scores	similar scores
SOM	13	4	3	0.128	-0.084	similar scores	similar scores
SPM	0	0	0	-0.063	-0.020	similar scores	similar scores
SRB	112	20	11	0.036	0.137	similar scores	similar scores
SSD	0	0	0	-0.063	-0.166	similar scores	lower in news
STP	0	0	0	-0.063	-0.020	similar scores	similar scores
SUR	3	1	1	-0.063	-0.137	similar scores	similar scores
SVK	89	7	4	0.178	-0.063	higher in news	similar scores
SVN	392	31	13	0.137	0.159	similar scores	higher in news
SWE	162	12	8	0.298	0.100	higher in news	similar scores
SWZ	0	0	0	-0.063	-0.067	similar scores	similar scores
SXM	0	0	0	-0.063	-0.020	similar scores	similar scores
SYC	13	6	2	0.057	0.101	similar scores	similar scores
SYR	6	3	2	0.057	-0.102	similar scores	similar scores
TCA	0	0	0	-0.063	-0.020	similar scores	similar scores
TCD	0	0	0	-0.063	-0.229	similar scores	lower in news
TGO	0	0	0	-0.063	-0.025	similar scores	similar scores
THA	177	31	18	-0.062	-0.039	similar scores	similar scores
TJK	12	6	6	-0.360	-0.301	lower in news	lower in news
TKL	0	0	0	-0.063	-0.020	similar scores	similar scores
TKM	0	0	0	-0.063	-0.298	similar scores	lower in news
TLS	61	5	1	-0.258	-0.206	lower in news	lower in news
TON	6	5	1	-0.063	-0.025	similar scores	similar scores
TTO	1	1	1	-0.309	-0.067	lower in news	similar scores
TUN	4	4	4	0.178	0.104	higher in news	similar scores
TUR	300	63	43	0.123	-0.021	similar scores	similar scores
TUV	0	0	0	-0.063	-0.020	similar scores	similar scores
TWN	530	84	47	0.094	0.023	similar scores	similar scores
TZA	84	25	15	0.063	0.108	similar scores	similar scores
UGA	236	47	36	0.012	0.402	similar scores	higher in news
UKR	4	4	4	0.178	-0.082	higher in news	similar scores
URY	11	3	3	0.128	0.166	similar scores	higher in news
USA	2,172	356	191	0.206	0.116	higher in news	similar scores
UZB	0	0	0	-0.186	-0.426	lower in news	lower in news
VAT	1	1	1	-0.063	-0.019	similar scores	similar scores
VCT	7	4	2	-0.137	0.101	similar scores	similar scores
VEN	294	54	33	0.137	0.097	similar scores	similar scores
VGB	0	0	0	-0.063	-0.020	similar scores	similar scores
VIR	0	0	0	-0.063	-0.020	similar scores	similar scores
VNM	673	96	54	-0.086	0.035	similar scores	similar scores
VUT	43	8	5	-0.101	-0.102	similar scores	similar scores
WLF	0	0	0	-0.063	-0.020	similar scores	similar scores
WSM	6	4	3	0.128	0.172	similar scores	higher in news
YEM	67	11	8	-0.046	-0.064	similar scores	similar scores
ZAF	406	46	28	0.271	0.192	higher in news	higher in news
ZMB	3	2	2	0.057	0.002	similar scores	similar scores
ZWE	32	6	4	0.055	0.104	similar scores	similar scores

Table 8: Computed measures for all countries.

CA-ClimateNews: A Human-Annotated Multilingual Benchmark for Climate-Relevance Classification of Central Asian News

Adilet Metinov[1] **Zarina Uvalieva**[1] **Bektemir Kumarbai uulu**[1]
Nurtilek Alibekov[1] **Tynchtykbek Tashbaltaev**[1]
[1]Airun Intelligence Lab, Bdigital LLC
metinovadilet@gmail.com

Abstract

Climate change is acute in Central Asia (glacier loss, the Aral Sea, and drought exposure), yet there are no NLP resources for climate discourse in the region's low-resource Turkic languages. We present **CA-ClimateNews**, the first human-annotated multilingual dataset for climate-relevance classification of Central Asian news, covering **Kyrgyz, Kazakh, Uzbek, and (regionally grounded) Russian**. The corpus comprises **3,140** news items labeled at two levels (binary climate relevance and an 8-way topic category) by two annotators per language, with a 100-item blind double-annotation overlap adjudicated to gold. Inter-annotator agreement is almost perfect (Cohen's κ 0.86–0.98 on relevance). Because we used transparent LLM-assisted pre-labeling, we additionally run a **between-subjects** experiment showing that **displaying model pre-labels induces anchoring**: an independent group shown the pre-label overrode it 3–7$\times$ less than a blind group. Our released labels are the blind version. We report baselines: fine-tuned multilingual encoders reach 88–96 relevance macro-F1 on the balanced benchmark; a Russian-only model loses up to $\sim$7 points transferring zero-shot to the Turkic languages *even at matched training size* (least, $\sim$2, for Kazakh), and a related Turkic source is no better, isolating a cross-lingual gap that in-language supervision closes. The dataset, guidelines, and code are publicly released.[1]

1 Introduction

Central Asia is on the front line of climate change: shrinking Tian Shan and Pamir glaciers (Farinotti et al., 2015) feed the region's rivers, the Aral Sea has largely disappeared (Micklin, 2007), and heat and drought threaten agriculture and water security (Shaw et al., 2022). Understanding how this is discussed in regional media requires being able to *find* and *categorize* climate content, a text-classification problem. Yet no annotated resources exist for climate discourse in the region's low-resource Turkic languages (Kyrgyz, Kazakh, Uzbek), and multilingual climate NLP has focused almost entirely on English (Diggelmann et al., 2020; Webersinke et al., 2021; Coan et al., 2021; Laud et al., 2023).

We close this gap with **CA-ClimateNews**, a human-annotated benchmark for two-level climate-relevance classification of Central Asian news in four languages. Building it surfaced a striking property of the domain: climate is a *needle in a haystack*: only $\sim$1% of a random news sample is climate relevant (1.00%; 58/5,780 LLM-labeled random items, 95% Wilson CI $[0.8, 1.3]$%), ranging from 0.34% (Kyrgyz) to 1.7% (Uzbek). This scarcity is precisely why the resource did not previously exist and why targeted collection is required.

Our contributions are:

1. **The first multilingual, human-annotated climate-relevance dataset for Central Asia** (3,140 items; Kyrgyz/Kazakh/Uzbek/Russian), with two-level labels, near-perfect inter-annotator agreement, and adjudicated gold.
2. **A between-subjects anchoring result**: an independent annotator group shown the LLM pre-label overrides it 3–7$\times$ less than a blind group, a concrete, quantified caution for LLM-assisted labeling pipelines.
3. **Baselines** revealing the low-resource gap: multilingual and monolingual fine-tuning, an LLM pre-labeling baseline, and an RU$\rightarrow$Turkic zero-shot transfer experiment with matched-data and Turkic$\rightarrow$Turkic controls, showing a cross-lingual drop of up to $\sim$7 points ($\sim$2 for Kazakh) that a related same-script source does not avoid and that subword overlap does not explain.

[1]https://github.com/ma-bdigital/CA-ClimateNews

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 20–28
October 28, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Climate NLP is overwhelmingly English: claim verification (Diggelmann et al., 2020), topic/relevance detection (Varini et al., 2020), environmental-claim detection (Stammbach et al., 2023), domain language models (Webersinke et al., 2021), contrarian-claim detection (Coan et al., 2021), and multi-task benchmarks (Laud et al., 2023). We found *no* climate or environmental text-classification resource for Kyrgyz, Kazakh, or Uzbek, a gap this paper addresses. The closest methodological cousin is MASAKHANEWS (Adelani et al., 2023), a multilingual human-annotated news-topic dataset for African languages whose pipeline, like many, had annotators correct classifier pre-labels, the practice we scrutinize in §6.

Central Asian NLP has grown but not toward climate. Kazakh has NER (Yeshpanov et al., 2022); Uzbek has news topic classification (Kuriyozov et al., 2023) and monolingual encoders (Mansurov and Mansurov, 2021; Kuriyozov et al., 2024); Kyrgyz has a multilabel topic benchmark (Alekseev et al., 2023) and, recently, a first monolingual encoder (Metinov et al., 2025); and Turkic understanding benchmarks exist (TUMLU (Isbarov et al., 2025), Kardeş-NLU (Senel et al., 2024), SIB-200 (Adelani et al., 2024)), with recent surveys mapping the landscape (Veitsman and Hartmann, 2025; Alekseev and Turatali, 2024), but none targets climate or news relevance. Cross-lingual coverage comes from multilingual encoders: XLM-R (Conneau et al., 2020), Glot500 (ImaniGooghari et al., 2023), mmBERT (Marone et al., 2025), though these over-fragment Turkic text (Rust et al., 2021), and zero-shot transfer (Conneau et al., 2018; Hu et al., 2020; Ruder et al., 2021) falters for low-resource, distant targets (Joshi et al., 2020; Senel et al., 2024), the regime our RU→Turkic experiment probes.

LLM-assisted annotation. LLMs increasingly serve as data annotators (Gilardi et al., 2023; Wang et al., 2021; Ding et al., 2023), so showing their pre-labels to human verifiers is now common practice. But pre-annotation is long known to bias annotators toward the suggestion (Fort and Sagot, 2010), an instance of automation bias and anchoring (Skitka et al., 2000; Tversky and Kahneman, 1974). We quantify this effect for LLM pre-labels (§6).

Lang	Items	%REL	Date range
RU	800	52.0	2009–2026
KK	740	49.6	2003–2026
KY	800	47.8	2003–2026
UZ	800	52.4	2015–2026
All	3,140	50.4	2003–2026

Table 1: Corpus overview. %REL = share labeled climate-relevant (L1). KK totals 740 rather than 800 because, after deduplication, its outlet archives yielded fewer distinct climate-relevant items to balance against.

3 Dataset

Sources and collection. We collected news items (headline, lead, URL, date) from nine public Central Asian outlets via RSS and archive sitemaps: Azattyk (KY), 24.kg (KY/RU), Egemen and Azattyq (KK), Kun.uz, Daryo.uz and Gazeta.uz (UZ), and Tengrinews, Kaktus and Sputnik (RU). Because climate content is rare (∼1% of random samples), the climate portion was collected by keyword/slug search over full outlet archives, and the non-relevant portion by random sampling. Crucially, the non-relevant half deliberately mixes *hard negatives*, climate-adjacent items (generic weather, pollution, or agriculture with no climate framing) drawn from the same keyword harvest, with easy random negatives, so a classifier cannot succeed by keyword presence alone. All items are deduplicated by normalized URL and normalized title. Both the keyword-seeded and random components are reported for transparency (§7). Language codes label the *article* language; the Russian portion is drawn from Central Asian outlets across all three countries, making it regionally grounded rather than Moscow-centric.

Size and balance. Table 1 summarizes the corpus: 3,140 items, balanced at 50.4% climate-relevant per the annotation design.

Label schema. Each item is labeled at two levels. **L1 (relevance, binary):** RELEVANT if substantively about climate change, its impacts, or responses; else NOT_RELEVANT (including generic weather or pollution with no climate framing). **L2 (topic, 8-way, only when L1=RELEVANT):** WATER_CRYO, DISASTER, AGRI_LAND, ENERGY, POLICY, SCIENCE_GEN, HEALTH_SOCIETY, OTHER_CLIMATE. Hard-case rules (weather/ disaster events relevant only if linked to climate; energy relevant only with renewable-transition/emissions/policy framing) are given verbatim in the guidelines; L2 takes the sin-

Lang	L1 κ	L1 %agr	L2 κ	L2 %agr
RU	0.859	93	0.826	86
KK	0.858	93	0.932	95
KY	0.860	93	1.000	100
UZ	0.980	99	0.872	90

Table 2: Inter-annotator agreement on the blind 100-item overlap. All Cohen's $\kappa \geq 0.81$ (almost perfect).

gle *dominant* frame, ties breaking to the more specific category. The final L2 distribution is dominated by WATER_CRYO (464) and POLICY (426), reflecting the region's glacier/water and policy-oriented coverage.

4 Annotation Protocol

Two native-speaker annotators labeled each language, following a shared guidelines document (in Russian, the common language) designed per standard annotation practice (Pustejovsky and Stubbs, 2012). The protocol was: (1) a 30-item Russian *calibration* round to align annotators; (2) a 100-item *blind double-annotation overlap* per language (LLM pre-labels hidden) for agreement; (3) *adjudication* of disagreements into gold; (4) blind single-annotation of the remaining items, split disjointly and source-balanced between the two annotators. Every item received a human decision; no item ships on an LLM label alone. Separately, for the anchoring experiment (§6), an *independent* group of two native speakers per language annotated the same items with the LLM pre-label visible; those labels are used only for that comparison and are not released.

5 Inter-Annotator Agreement

On the 100-item blind overlap per language, Cohen's κ (Cohen, 1960) is in the "almost perfect" band (Landis and Koch, 1977) for both levels (Table 2); see Artstein and Poesio (2008) on agreement measures. L2 κ is computed only on the items both annotators called relevant ($42/40/47/52$ for RU/KK/KY/UZ), so KY's $\kappa = 1.000$ is perfect agreement on 47 items. The 35 disagreements were adjudicated into gold, each decision released per item. We further **triple-annotate the held-out test set**: two more annotators re-labeled all 630 items *blind* (three-way L1 κ 0.75–0.99, 90.3% unanimous; App. B), and majority adjudication yields the released *test gold* used for all baseline metrics.

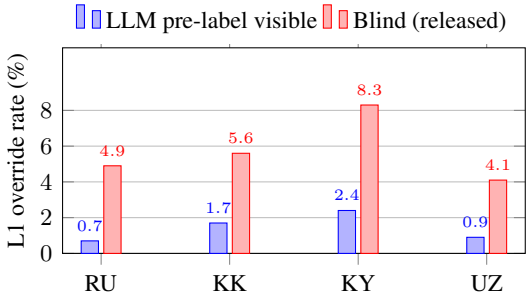


Figure 1: L1 override rate for an *LLM-visible* annotator group vs. an independent *blind* group (two native speakers per language each). The blind group overrides 3–7× more often, exposing anchoring; its rate approaches the LLM-error estimate (2–14%) and the human–human floor (1–7%).

6 The LLM Pre-label Anchoring Effect

We used transparent LLM-assisted pre-labeling: a large language model (Anthropic's Claude Sonnet 5), prompted with the annotation guidelines at default decoding, proposed L1/L2 labels for humans to verify or correct; the exact prompt and model version are released with the code. To test whether the pre-label biases annotators, we ran a **between-subjects** experiment: for each language, one group of two native speakers annotated with the pre-label *visible* (and were told it was model-generated), and a *separate, independent* group of two native speakers annotated the same items *blind* (pre-label hidden). Because the groups are disjoint, the comparison is free of the within-annotator ordering, learning, and memory confounds that a single-group before/after design would incur.

Each item was single-annotated; an override is any human label that differs from the model's. Over the full single-annotated bulk (N=2,740; 640–700/language, *not* the overlap), the blind group overrode the model on 157 items vs. the visible group's 39 (5.7% vs. 1.4%; 118 more corrections), and the blind rate exceeds the visible rate in *all four* languages by a large margin (3–7×; mean +4.3 points; Figure 1). Because each condition has only two annotators, an item-level test would overstate significance. But the conditions use *disjoint* annotator sets (sixteen people, none in both), so the $4/4$ replication is across distinct individuals, not just languages (App. C). The comparison is against the model's *fixed* pre-labels, so it does not depend on gold, and the blind rate approaches the LLM-error and human–human floors (Figure 1). **Recommendation:** hide model pre-labels, or elicit

Lang	L1 Acc	L1 κ	L2 Acc	L2 κ
RU	94.8	0.895	85.5	0.822
KK	93.2	0.865	89.8	0.852
KY	91.5	0.831	85.1	0.795
UZ	96.1	0.922	78.0	0.727
All	93.9	0.878	84.4	0.802

Table 3: LLM pre-label baseline vs. human labels (% and κ). L2 on items both call relevant.

the human label first; our released labels are the blind version.

7 Baselines

LLM pre-labeling. Against the human labels, the LLM pre-labeler reaches 93.9% L1 accuracy (κ 0.878) with high recall (98.5%) but lower precision (90.3%): it over-includes, functioning as a strong recall-oriented pre-filter that still needs human review (Table 3). L2 accuracy is 84.4% (κ 0.802); the top confusions (OTHER_CLIMATE↔SCIENCE_GEN, WATER_CRYO↔DISASTER) coincide with the categories the two humans disagreed on: genuine frame ambiguity, not random error.

Supervised and transfer baselines. These *characterize* the benchmark; they are untuned reference points, not tuned systems. We fine-tune encoders on a 70/10/20 split stratified by language × (L1,L2), and report per-language macro-F1 on the held-out test set (Table 4). We compare: (i) a majority-class floor; (ii) massively multilingual encoders: XLM-R (Conneau et al., 2020), Glot500 (Imani-Googhari et al., 2023), mmBERT (Marone et al., 2025), mBERT (Devlin et al., 2019), fine-tuned jointly on all four languages; (iii) per-language monolingual encoders (Kaz-RoBERTa (Sagyndyk et al., 2025), TahrirchiBERT, KyrgyzBERT2 (the latest release of Metinov et al., 2025), RuBERT); and (iv) **cross-lingual transfer**: XLM-R fine-tuned on Russian only and evaluated *zero-shot* on the three Turkic languages, with a matched-data *in-language control* (XLM-R on the same ∼560 target items) and a *Turkic→Turkic control* (KK→KY) that together separate cross-lingual loss from data quantity and from source distance (Table 5). Main baselines and the KK→KY control report the mean over three seeds (13/42/100); the RU→Turkic and in-language rows use eight.

Results. On the balanced (50/50) test set, all multilingual encoders reach 88–96 L1 macro-F1, far above the ∼33 majority floor; no single encoder dominates (best multilingual: Glot500 on KY, XLM-R on KK/RU, mmBERT on UZ). No monolingual clearly beats them (Kaz-RoBERTa is level on Kazakh within seed variance; the Kyrgyz/Uzbek encoders trail ∼6/∼14; Table 4). Our main characterization result is the **RU→Turkic transfer gap** (Table 5): at a matched budget (∼560 items, 8 seeds), zero-shot transfer from Russian costs 2.0–6.6 macro-F1 points versus in-language training (smallest for Kazakh). This is *not* straightforwardly explained by script or lexical distance: a same-family Turkic source transfers no better (KK→KY 86.2; Table 5), and subword overlap does not track the gap, reflecting the value of *in-language* supervision that no related source substitutes. L2 (8-way topic) is harder and noisier (best mmBERT 0.58–0.75; Appendix A). We release CA-ClimateNews (near-perfect agreement, adjudicated gold, transparent LLM assistance, an anchor-

Model	KY	KK	UZ	RU
Majority	30.7	32.6	34.8	34.7
Multilingual (joint fine-tune)				
XLM-R	90.9	87.9	89.6	**96.2**
Glot500	**93.2**	87.7	91.7	95.2
mBERT	91.1	87.5	90.6	95.4
mmBERT	91.9	84.1	**94.6**	94.8
Best mono.	87.7	**88.3**	81.1	92.3

Table 4: Level-1 (climate relevance) macro-F1 (×100), mean over 3 seeds (13/42/100); per-cell std 0.4–3.3 (largest for mBERT on KK). Best per language in **bold**. Best mono. = KyrgyzBERT2 / Kaz-RoBERTa / TahrirchiBERT / RuBERT, each fine-tuned on its own language. Monolingual Kaz-RoBERTa edges the best multilingual on KK by 0.4 (within seed variance); elsewhere the multilingual encoders lead by ∼4 on RU and 5–14 on KY/UZ.

XLM-R training setup	KY	KK	UZ
In-language (mono, L-only)	91.3	88.3	83.7
RU-only → zero-shot	86.8	86.3	77.1
KK-only → zero-shot†	86.2	–	–
Δ (in-lang − RU-only)	+4.5	+2.0	+6.6

Table 5: RU→Turkic transfer at a matched ∼560-item budget (Level-1 macro-F1 ×100, mean over 8 seeds; std ≤ 4.0). These matched-budget in-language numbers use only ∼560 items, so fall below the full-data joint fine-tune of Table 4 (notably UZ). Zero-shot from Russian costs 2.0–6.6 points; the Kyrgyz/Uzbek gaps are positive in all 8 seeds, the Kazakh gap (+2.0) within seed variance. †Turkic→Turkic control (KK→KY), mean over 3 seeds.

ing caution) with baselines showing a RU→Turkic gap that in-language supervision closes.

Limitations

The held-out *test set* is triple-annotated and the 400 overlap items are gold (§5); the remaining training items are single-annotated blind. Human–human and LLM-error rates are estimated on 100 items/language (sampling noise; the anchoring *direction* is robust across all four languages). The corpus is news register only, uses a headline+lead unit (the climate frame is set in the lead), and a single-label 8-way scheme; we make no dialectal-coverage claims. Collection mixes keyword-seeded (climate) and random (non-climate) sampling, and includes state-affiliated outlets; both are disclosed. Because the climate-relevant items are keyword/slug-retrieved, climate coverage that avoids the seed vocabulary is under-represented in the positive class (the hard-negative design guards the negative class but not this); the benchmark therefore measures relevance discrimination rather than open-world recall. Some items lack a parsed date (mostly one outlet); date ranges differ by language (KY and KK reach back to 2003, RU to 2009, UZ only to 2015) because of archive depth, so outlet and period are confounded with language. The test set is balanced (∼50% relevant) for stable macro-F1; at the ∼1% deployment prevalence the models are best read as a high-recall *retrieval* filter, not final classifiers. The low resulting precision is the base-rate effect, not a model flaw: at ∼90% recall, prevalence-adjusted precision ranges from ∼4% (Kyrgyz) to 31% (Russian) (Table 9), yet the triage value is large: a reviewer then inspects only 4–15% of the stream (7–26× fewer items), and raising the threshold to 80% recall lifts precision to 52–100% on Russian/Uzbek. The prevalence estimates (∼1% overall; ∼1,400–1,500 random items/language) rest on few positives (as few as five for Kyrgyz), so per-language values carry wide confidence intervals and Table 9's deployment numbers are illustrative.

Ethics and Data Statement

We release headlines, leads, URLs, access dates, and labels, not full article text. The **annotations/labels are released under CC BY 4.0** and the **code under the MIT License**; the headline and lead excerpts remain the property of their respective outlets and are provided as short excerpts under fair-use/quotation norms, with a takedown process for any outlet that requests it. All labels are keyed to source URLs and access dates, so the text can be re-fetched from source if preferred. The study involved four native-speaker annotators per language (two who produced the released blind labels and two independent annotators for the LLM-visible anchoring condition), plus two further annotators who blind-re-labeled the held-out test set, all university-educated, aged 20–30, residing in the respective countries, and compensated at local professional rates. Known biases: news register only, a keyword-seeded climate portion, and state-affiliated outlets in the mix.

Acknowledgments

We thank the native-speaker annotators for Kyrgyz, Kazakh, Uzbek, and Russian for their careful work, and the anonymous reviewers for their helpful feedback.

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, et al. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of IJCNLP-AACL (Volume 1: Long Papers)*.

David Ifeoluwa Adelani et al. 2024. SIB-200: A simple, inclusive, and big benchmark for topic classification in 200+ languages and dialects. In *Proceedings of EACL*. ArXiv:2309.07445.

Anton Alekseev, Sergey I. Nikolenko, and Gulnara Kabaeva. 2023. Benchmarking multilabel topic classification in the Kyrgyz language. *arXiv preprint arXiv:2308.15952*.

Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: Challenges, progress, and future. *arXiv preprint arXiv:2411.05503*.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of ACL*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of ACL (Volume 1: Long Papers)*, pages 11173–11195.

Daniel Farinotti, Laurent Longuevergne, Geir Moholdt, Doris Duethmann, Thomas Mölg, Tobias Bolch, Sergiy Vorogushyn, and Andreas Güntner. 2015. Substantial glacier mass loss in the Tien Shan over the past 50 years. *Nature Geoscience*, 8(9):716–722.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pages 56–63.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, volume 119, pages 4411–4421.

Ayyoob ImaniGooghari et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of ACL*. ArXiv:2305.12182.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, et al. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. *arXiv preprint arXiv:2502.11020*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*, pages 6282–6293.

Elmurod Kuriyozov, David Vilares, and Carlos Gómez-Rodríguez. 2024. BERTbek: A pretrained language model for Uzbek. In *Proceedings of SIGUL*.

Elmurod Kuriyozov et al. 2023. Text classification dataset and analysis for Uzbek language. *arXiv preprint arXiv:2302.14494*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Tanmay Laud, Daniel Spokoyny, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. ClimaBench: A benchmark dataset for climate change text understanding in english. In *arXiv preprint arXiv:2301.04253*.

B. Mansurov and A. Mansurov. 2021. UzBERT: pretraining a BERT model for Uzbek. *arXiv preprint arXiv:2108.09814*.

Marc Marone et al. 2025. mmBERT: A modern multilingual encoder. *arXiv preprint arXiv:2509.06888*.

Adilet Metinov, A. Kudakeeva, and A. Kabaeva. 2025. KyrgyzBERT: A compact, efficient language model for Kyrgyz. *arXiv preprint arXiv:2511.20182*.

Philip Micklin. 2007. The Aral Sea disaster. *Annual Review of Earth and Planetary Sciences*, 35:47–72.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP*, pages 10215–10245.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP (Volume 1: Long Papers)*, pages 3118–3135.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, and Kirill Yakunin. 2025. Kaz-RoBERTa conversational: Technical report. *TechRxiv*. DOI 10.36227/techrxiv.175942902.25827042/v1.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schütze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of EACL (Volume 1: Long Papers)*, pages 1672–1688.

R. Shaw, Y. Luo, T. S. Cheong, S. Abdul Halim, S. Chaturvedi, M. Hashizume, G. E. Insarov, Y. Ishikawa, M. Jafari, A. Kitoh, J. Pulhin, C. Singh, K. Vasant, and Z. Zhang. 2022. Asia. In H.-O. Pörtner et al., editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the IPCC*, pages 1457–1579. Cambridge University Press.

Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Automation bias and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, 10(1).

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 1051–1066.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157).

Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. ClimaText: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. ArXiv:2012.00483.

Yana Veitsman and Mareike Hartmann. 2025. Recent advancements and challenges of Turkic Central Asian language processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM)*, pages 309–324.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. ClimateBERT: A pretrained language model for climate-related text. In *arXiv preprint arXiv:2110.12010*.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of LREC*. ArXiv:2111.13419.

Model	KY	KK	UZ	RU
Majority	9.4	8.9	6.4	5.5
XLM-R	58.0	62.5	49.1	48.3
Glot500	29.2	28.8	35.3	32.2
mBERT	55.6	63.8	48.3	52.9
mmBERT	**70.2**	**75.3**	**65.4**	**58.2**
Best mono.	41.1	37.8	22.0	36.3

Table 6: Level-2 (8-way topic) macro-F1 ($\times 100$) on relevant test items, mean over 3 seeds (13/42/100). mmBERT dominates; Glot500 collapses and is high-variance (std up to 13) on this fine-grained, small-sample task.

Lang	A–B	A–M	B–M	unan.	flips
RU	0.900	0.875	0.875	91.2%	6
KK	0.840	0.852	0.799	87.2%	7
KY	0.869	0.749	0.799	85.4%	13
UZ	0.988	0.938	0.951	96.9%	4
All	0.901	0.854	0.857	90.3%	30

Table 7: Test-set Level-1 agreement across three independent annotators. A, B = the two new blind annotators; M = existing master label. unan. = three-way unanimity; flips = L1 labels changed by majority adjudication.

A Level-2 Topic Results

Table 6 reports per-language macro-F1 for the 8-way topic task (L2), evaluated on the relevant items of each language's test set ($\sim$75–90 items). Because several categories are rare, macro-F1 is noisy; treat these as indicative. mmBERT is strongest overall. Glot500's collapse (L2 $\sim$30 vs. L1 $\sim$91) is a training-instability artifact, not a representational limit: its L2 macro-F1 varied widely across seeds (std up to 13, vs. ≤ 5 for the other multilingual encoders); we report it unchanged for transparency. The small monolingual models also struggle on this fine-grained task.

B Test-Set Triple Annotation

Two annotators independently re-labeled the entire held-out test set *blind*; together with the existing label this gives three independent labels per item. Table 7 reports pairwise Level-1 Cohen's κ (A, B = the two new annotators; M = the existing master label), three-way unanimity, and the number of labels changed by majority adjudication. All pairwise κ are in the "almost perfect" band; majority voting corrects 30 of 630 L1 labels (4.8%), which become the released test gold used for all Section 7 metrics.

Lang	N	vis.	blind	vis.%	blind%	$\times$
RU	700	5	34	0.7	4.9	6.8
KK	640	11	36	1.7	5.6	3.3
KY	700	17	58	2.4	8.3	3.4
UZ	700	6	29	0.9	4.1	4.8

Table 8: Anchoring override counts by condition over the single-annotated bulk. vis./blind = items (and %) where the human label differed from the model in the LLM-visible / blind condition; $\times$ = blind/visible ratio. We report effect sizes and four-way replication rather than an item-level p-value (see text).

C Anchoring Significance

The anchoring override rates (§6, Figure 1) are measured over the full single-annotated bulk, not the 100-item overlap. Table 8 gives the per-language item counts N, the visible- and blind-condition override counts and rates, and the blind/visible ratio. Because each condition has only two annotators who split the items disjointly, override decisions are not independent within a condition, so an item-level test (e.g. a two-proportion z treating 2,740 items as independent) would badly overstate significance. We therefore analyze at the language level: the blind rate exceeds the visible rate in all four languages by a large, consistent margin (mean $+4.3$ points, 3–7$\times$; one-sided sign test p=0.06). Because the two conditions use disjoint annotator sets (sixteen people, none in both), this four-way replication is across distinct individuals; we rest the anchoring claim on the effect size and replication rather than on a single small-sample p-value.

D Deployment Operating Points

Balanced macro-F1 measures discrimination; deployment on the natural $\sim$1% climate stream is a retrieval task. Table 9 reports, for each language's best encoder, the precision–recall tradeoff *at the natural prevalence*: at each recall target, the fraction of the stream flagged (a reviewer's reading load), the prevalence-adjusted precision, and the resulting load reduction. Precision is tunable: raising the threshold to 80% recall lifts precision substantially (to 100% on Russian, 52% on Uzbek) while flagging under 3% of the stream. Kazakh and Kyrgyz stay low-precision even then, reflecting higher false-positive rates.

Lang	model	recall	flagged	prec	load
RU	XLM-R	90%	3.8%	31%	26×
RU	XLM-R	80%	1.0%	100%	96×
KK	XLM-R	90%	14.5%	5%	7×
KK	XLM-R	80%	7.5%	8%	13×
KY	Glot500	90%	7.4%	4%	14×
KY	Glot500	80%	1.7%	16%	59×
UZ	mmBERT	90%	5.4%	28%	19×
UZ	mmBERT	80%	2.7%	52%	37×

Table 9: Deployment operating points at the natural per-language prevalence (KY 0.34%, KK 0.74%, UZ 1.69%, RU 1.30%). flagged = share of the stream predicted relevant (reading load); prec = prevalence-adjusted precision; load = reading-load reduction ($1/\text{flagged}$) vs. reading everything.

Climatextract: A Systematic Prompt Evaluation for Climate Information Extraction

Anna Steinberg Schulten[1,5,*], **Malte Schierholz**[1,5], **Lisa Reichenbach**[2], **Frauke Kreuter**[1,4,5], **Alexander Fraser**[3,5]

[1]LMU Munich,
[2]Deutsche Bundesbank, [3]Technical University Munich
[4]University of Maryland, [5]Munich Center for Machine Learning (MCML)

Abstract

Automated extraction of corporate greenhouse gas (GHG) emissions from sustainability reports is critical for climate crisis governance, yet the choice of prompt design, context provision, and model in LLM-based pipelines is largely guided by informal best-practice recommendations rather than systematic evaluation. With **Climatextract**, we assess these choices in a RAG pipeline, comparing closed-source against open-weight models, a structured "best-practice" prompt against a simple question-style prompt, and text-only against text-plus-table contexts. We find that performance is non-monotonic across model generations, a structured prompt with roles, definitions, extraction rules, and JSON formatting is significantly outperformed by a question-style prompt, and tabular context helps only in a specific model-prompt combination. A cost-benefit analysis shows that performance gains are often disproportionate to cost increases.

1 Introduction

Consistent and reliable data on company-level greenhouse gas (GHG) emissions are essential for effective climate policy, financial risk assessment, and sustainability research (Network for Greening the Financial System, 2024). In practice, however, due to a lacking mandatory reporting format, companies typically report their GHG emissions in non-standardized, unstructured, and lengthy PDF documents and upload them to their individual websites, rather than through a centralized repository. Scalable and transparent systems for automatic extraction of these values have thus been increasingly studied by the ClimateNLP community (Schimanski et al., 2024; Mankour et al., 2025; Forster et al., 2025).

Despite widespread deployment of LLMs for this task, the choice of prompt design, context provision, and model remains largely guided by informal best-practice recommendations from online websites (e.g. Prompting Guide from DAIR.AI) and model providers, or by ad hoc decisions reported without systematic evaluation. Prior work either reports only the best-performing prompt for a specific use case (e.g. Neuberger et al., 2025) or evaluates a single prompting technique like prompt chaining in isolation (Kwak et al., 2024). This raises the question: do the widely recommended prompt design choices actually improve emissions extraction from corporate sustainability reports?

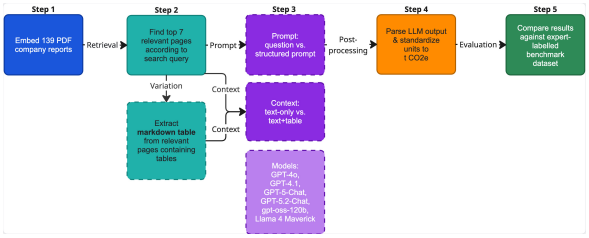


Figure 1: Overview of the **Climatextract** pipeline for extracting firm-level GHG emissions. Report pages are embedded and the most relevant ones retrieved, optionally with table extraction. An LLM is then prompted with the retrieved context, and its output is parsed and evaluated against the benchmark dataset. Dashed boxes mark the components varied in our systematic evaluation (context type, prompt template, model).

We address this question with **Climatextract**, a tool to perform systematic evaluation of prompt design choices for LLM-based GHG emissions extraction. Building on a retrieval-augmented generation (RAG) pipeline (Figure 1), similar to Mankour et al. (2025) and Forster et al. (2025), but we focus primarily on the LLM-based generation step (as opposed to retrieval) and assess three factors: model choice (four gpt generations and two open-weight models), prompt structure (a structured prompt following best practices vs. a simple question-style prompt), and context type (text-only vs. text with Markdown-serialized tables).

* Corresponding author: anna.steinberg@lmu.de

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 29–40
October 28, 2026 ©2026 Association for Computational Linguistics

Our experiments show that design choices commonly recommended in online prompt-engineering guides, including structured prompt templates, table-aware context, and the newest available model, do not consistently translate into stronger extraction quality. Model release order is an unreliable guide to performance, and the open-weight models we test trail meaningfully behind their closed-source counterparts. A structured prompt built from a role, definitions, extraction rules, and JSON formatting is beaten by a simple question-style variant. The value of adding tabular context turns out to hinge on the specific model and prompt used. Cost differences between configurations are also often out of proportion to the corresponding gains in extraction quality. We release our pipeline as a Python package on PyPI[1] to support systematic prompt evaluation and GHG emission extraction on new climate corpora. All experiments are reproducible from the publicly available benchmark dataset (Beck et al., 2025).

2 Related Work

A common pattern across recent work on LLM-based extraction from sustainability reports is to rely on RAG pipelines and reformulate extraction tasks as question-answering problems, operating primarily on plain-text representations of the underlying PDFs (Ni et al., 2023; Vaghefi et al., 2023; Schimanski et al., 2024; Dimmelmeier et al., 2024). A recurring finding is that retrieval quality critically shapes extraction accuracy, while attempts to preserve document or tabular structure in embeddings often reduce performance (Mankour et al., 2025). Recent work has moved toward large-scale, end-to-end extraction pipelines capable of handling hundreds of indicators, though most offer limited transparency on prompt design, invoking general "best practices" rather than empirically testing prompt components (Forster et al., 2025).

Since companies often report emissions data in tables, table handling is increasingly treated as a dedicated subprocess: Ali et al. (2025), for example, use an LLM to summarize tables into paragraphs before generating question-answer pairs, whereas we retain the tabular structure by converting raw text into Markdown tables and providing them as additional context, following the recommendation of Yue et al. (2025).

The closest analogues to our work are Atreja et al. (2025) and Guo et al. (2024), who systematically vary model selection and design features and consolidate their findings into evidence-based recommendations. Atreja et al. (2025) vary definitions, output type, and prompt length; Guo et al. (2024) vary input, output, and reasoning design for fine-tuning samples. Their results, however, reveal inconsistencies across models and tasks. We build on this line of work and investigate similar variations for prompt-based emissions extraction from sustainability reports.

3 Method

Our pipeline proceeds in 5 steps as displayed by Figure 1.

Step 1: Document Embedding. The pipeline takes raw PDF reports as input. In the first step, we extract the text of each PDF page via `PDFReader` and embed each page separately using OpenAI's `text-embedding-ada-002` (Greene et al., 2022). Pages are loaded via `PDFReader` and treated as atomic chunks, i.e. no sub-page or overlapping chunking is applied.[2]

Step 2: Retrieval. We then identify relevant context by semantically searching for pages most relevant to the query: "What are the total CO2 emissions in different years? Include Scope 1, Scope 2, and Scope 3 emissions if available.". Scope 1, 2 and 3 refer to different operational boundaries as they are specified by the reporting guidelines of the Greenhouse Gas Protocol (GHGP) (WRI and WBCSD, 2011)[3]. We then consider the top 7 pages in terms of cosine similarity to the query as relevant context for the prompting step. Prior empirical evaluations (see Appendix A.1) indicated that the number of retrieved pages set to 7 represents the best trade-off between accuracy and computational cost for our use case. Since the retrieval query and top-k parameter are held constant across configurations, each report yields the same retrieved pages for every downstream comparison. Retrieval nevertheless remains part of the evaluated pipeline

[1] https://pypi.org/project/climatextract/0.3.2/

[2] Pages exceeding the embedding model's 8,192-token limit are truncated only for the embedding vector; the full page text is passed to the LLM at extraction time if that page is retrieved.

[3] Scope 1 emissions refer to direct greenhouse gas emissions from sources that are owned or controlled by the company. Scope 2 covers indirect GHG emissions arising from the purchase and use of electricity, disclosed as either according to location- or market-based methods (2lb and 2mb, respectively). Scope 3 emissions cover 15 diverse categories of residual up- and downstream value-chain emissions.

because retrieval failures, namely cases where relevant pages are not among the top-k retrieved, do occur for 2 of 139 reports in our benchmark (see Figure A4 in the appendix). Presenting results on pre-computed static context would eliminate this failure mode and inflate performance.

Optional table extraction. The raw text extracted from PDF documents does not capture the tabular layout, leading among other issues to potential misattribution of values to incorrect years (e.g., a value for 2015 might be linked to 2016). Given that 85% of our ground-truth values are entered in tables, we follow Ali et al. (2025) to process these separately by checking if relevant pages contain tables and converting the filtered pages to markdown format to preserve structure.

Step 3: Prompting. We vary three factors:

- Model: To assess the trade-off between cost and performance, we test recent closed-source models (gpt-4o (OpenAI, 2024), gpt-4.1 (OpenAI, 2025b), gpt-5 (Singh et al., 2025) and gpt-5.2 (OpenAI, 2025c) against two open-weight models, gpt-oss-120b (OpenAI, 2025a) and Llama 4 (Meta AI, 2025). To minimise stochastic variation, we set temperature to 0 for all models. Since the task is pure information extraction and does not involve logical or mathematical thinking, we consider the chat versions of gpt-5 and gpt-5.2 and do not consider any reasoning models for these experiments, with the exception of gpt-oss-120b where we set `reasoning_effort` to `"none"` to disable reasoning behaviour.

- Prompt structure: We compare the *question prompt* against the *structured prompt*. The *question prompt* (see Figure A1 in the appendix) states the task at the beginning, includes an instruction for the case of no values found and then lists all 44 combinations of 4 metrics (Scope 1, 2mb, 2lb and 3) and 11 years of interest as questions, appends the context and an example of the preferred response format, constructed by reiterating the first 3 questions and using placeholders. Following best practices, the *structured prompt* (see Figure A2 in the appendix) defines a role for the LLM, provides clear instructions, generates one-sentence long definitions of the scopes and lists the extraction rules reflecting the relevance criteria described below. For instance, the model is supposed to extract only absolute values, not percentage increases or emission intensities.

- Context type: We compare passing only textual context of the relevant pages to passing textual and markdown converted tabular context of relevant pages.

Following the benchmark's annotation schema, we aim to extract disclosed values only if they (i) refer to the whole company, (ii) are reported according to the operational boundaries of the scopes, (iii) are reported in absolute terms as CO2 or CO2e and (iv) represent a total value, not subcategories.

Step 4: Post-processing. We parse the LLM output, via regex for the question prompt and JSON parsing for the structured prompt, normalise units using a unit dictionary (standardising to tonnes of CO2 equivalents (CO2e)) and removing duplicates via custom prioritization rules (see Appendix A.3 for details).

Step 5: Evaluation. We compare extracted values to the ground truth values from the benchmark dataset and compute report-level precision, recall and F1-score. We rely on a publicly available benchmark, expert-annotated dataset of sustainability information from a sample of 139 sustainability and annual reports covering the year range 2017 to 2022, where each entry represents a combination of report, year and scope (Beck et al., 2025). The sector- and region-diverse set of companies publishing the PDF reports are a random selection of around 4000 companies featured in two major indices (international "MSCI World Small Cap" and German "DAX").

4 Results

4.1 Model experiments

Among the tested models (Table 1), gpt-5.2 delivers the strongest overall performance. Progress across gpt generations is uneven, however: while gpt-4.1 improves on gpt-4o across all three metrics, gpt-5 slips behind gpt-4.1 on precision and F1 despite reaching the highest recall of the closed-source models. The open-weight models fall considerably short of the gpt family. Notably, Llama-4 despite having more than twice the number of parameters of gpt-oss-120b (400B vs. 117B) shows highly unstable recall and ends up performing worse than the smaller gpt-oss-120b overall. Formal significance tests (indicated by stars in Table 1) confirm that gpt-5.2 significantly outperforms gpt-5, Llama-4, and

Model	Created	Precision	Recall	F1 Score
gpt-4o	11/24	0.656 (0.007)	0.929 (0.005)	0.769 (0.006)
gpt-4.1	04/25	0.658 (0.017)	0.944 (0.017)	0.775 (0.009)
gpt-5	08/25	0.644 (0.009)	**0.950** (0.008)	0.768** (0.007)
gpt-5.2	12/25	0.779 (0.015)	0.878 (0.084)	**0.823** (0.039)
Llama-4	04/25	0.182 (0.034)	0.817 (0.159)	0.293*** (0.043)
gpt-oss	08/25	**0.812** (0.040)	0.281 (0.042)	0.416*** (0.047)

Table 1: Averages and standard deviations of 5 runs of model experiments. The closed-source models show a nonmonotonic increase in performance (gpt-4.1 better than newer gpt-5, newest gpt-5.2 leaping ahead). The open-weight models lag behind the closed-source models. Significance markers on F1 scores indicate one-sided hypothesis tests against the reference model (gpt-5.2). Details in Appendix A.4. ***p<0.01, **p<0.05, *p<0.1.

Prompt Type	Input Mode	Precision	Recall	F1 Score
question	text	**0.779** (0.015)	0.878 (0.084)	0.823 (0.039)
question	text+table	0.777 (0.021)	**0.903** (0.014)	**0.835** (0.017)
structured	text	0.760 (0.008)	0.860 (0.011)	0.807 (0.006)
structured	text+table	0.755 (0.020)	0.830 (0.105)	0.787 (0.049)

Table 2: Averages and standard deviations of 5 runs of prompt and context experiments with gpt-5.2. A simple question-style prompt outperforms the structured prompt. Adding tabular context improves results for question prompt slightly, but not consistently for structured prompt.

gpt-oss, but the differences to gpt-4o and gpt-4.1 are not statistically supported (see Appendix A.4).

What the model extracts, mirrors what companies report: Scope 1 is most often extracted, Scope 2 location-based exceeds market-based since it is the default per our instruction if Scope 2 is not explicitly disclosed, and Scope 3 is extracted least often given its disclosure is not mandatory (cf. Figure A5).

A closer inspection reveals that gpt-oss-120b's low score is substantially a parsing artifact rather than a capability gap: 85.8% of its errors on non-NA reports stem from model-specific output formatting (e.g., bare numbered lists, Unicode punctuation substitution, missing colon separators) that our regex parser cannot handle, rather than incorrect values. Pairing gpt-oss-120b with structured JSON output instead of the regex-parsed question prompt raises its F1 from 0.416 to 0.786 (5-run average), on par with gpt-4.1 (see Appendix A.5 for details).

4.2 Prompt experiments

Using gpt-5.2, our best-performing model, we next compare the two prompt designs (Table 2). Contrary to best practice, the simple question prompt (row 1) achieves higher precision, higher recall, and consequently a higher F1 score than the structured prompt (row 3); the F1 difference is statistically significant (one-sided $p < 0.05$; see Appendix A.4).

Prompt	Precision	Recall	F1	ΔF1
Full structured	0.760 (0.008)	0.860 (0.011)	0.807 (0.006)	—
– role	0.823 (0.008)	0.819 (0.013)	0.821 (0.006)	+0.014
– definitions	0.790 (0.031)	0.827 (0.013)	0.808 (0.017)	+0.001
– rules	0.634 (0.045)	0.726 (0.026)	0.676 (0.030)	**–0.131**

Table 3: Averages and standard deviations of 5 runs of ablations of structured prompt components, ordered by effect on F1. Removing extraction rules causes a substantial performance drop, while removing the role slightly improves performance and removing definitions has no measurable effect.

To identify which parts of the structured prompt account for this outcome, we ablate its components individually (Table 3). Their contributions differ sharply. Dropping the role description nudges F1 slightly upward, from 0.807 to 0.821; dropping the scope definitions has essentially no effect while dropping the extraction rules causes F1 to plummet. The strong effect of the extraction rules is worth highlighting, given that a qualitative inspection of extracted values shows that the model occasionally ignores them anyway[4]. Even the strongest ablation variant, however, still falls short of the question prompt's F1, indicating that the structured format as a whole, not any single omitted component, drives the underperformance.

The comparison between question and structured prompts is not fully controlled: the two designs differ not only in output format but also in preamble content and length, and prior work has shown that requiring structured output (Tam et al., 2024) and increasing distance between task definition and context (Liu et al., 2024) can affect performance. Our ablation of individual preamble components (Table 3) partially addresses this by showing that removing the role description or scope definitions does not close the gap; a further ablation pairing the question prompt with JSON output (see Appendix A.6) also reduces rather than improves its F1, though with overlapping confidence intervals. A fully controlled comparison isolating output format from preamble structure remains an open direction. We thus interpret the finding as evidence that the structured prompt as a whole underperforms

[4]For instance, the pipeline extracts a value for a subcategory of Scope 3 instead of the total.

the question prompt in this setting, rather than as an isolated claim about JSON output format.

Categorising each report's extraction result under both prompts (see Figure A4 in the appendix) shows that more than 85% remain in the same category (correct, partially correct, incorrect) while the prompt design shifts only a handful of borderline cases. Two systematic error patterns appear under both prompts: value–year/scope misalignment, and extraction of subcategory values when both subcategories and totals for the same scope-year appear on the same page. Overall, the model handles reports without ground-truth values better than reports with them (91% correctly returning nothing vs. 51% correctly extracting the disclosed values).

4.3 Tabular context experiments

Adding the Markdown-serialized table as context yields a modest F1 gain for the question prompt, primarily through additional extractions (Figure A5) leading to improved recall (rows 1 and 2 of Table 2); this improvement is statistically significant (one-sided $p < 0.01$; see Appendix A.4).

For the structured prompt, the same intervention has the opposite effect, reducing both precision and recall. To probe how robust this pattern is, we repeat the experiments with our second-best model, gpt-4.1 (Table A5); here, tabular context degrades performance for both prompt types.[5]

Manual inspection of the 28 erroneous text-only extractions with ground-truth values in tables clarifies the limited effect: only 6 exhibit resolvable misalignment errors, and just 2 are corrected by the added context. The remaining 22 reports exhibit errors, unaddressed by tabular structure, such as irrelevant extractions or footnote misinterpretations.

All results reported in this paper are based on raw values. Inspecting the unit normalization as part of the post-processing in more detail, we find that text+table context increases the rate at which extracted units fail to resolve to a known category and instead fall into *Other* (Table A6), a discrepancy not reflected in the raw-value metrics reported above. Examining a representative report (Paramount Group, 2020, page 47), the unit "metric tons CO2e" appears only once, in the table header, and not alongside the individual Scope 1 values themselves. Under text-only context, the model infers this header unit and appends it to each extraction; under text+table context, it omits the "e" suffix and appends "metric tons CO2" which is absent from the unit mapping. Notably, the markdown table only contains values and no units.

The unresolved unit rate is also uneven across scopes, with Scope 1 and Scope 2lb tracking closely and Scope 3 remaining lowest throughout (see Appendix A.9 for details).

[5] Since the Markdown conversion is identical across prompt conditions, any potential conversion errors would affect them symmetrically and cannot account for the diverging results.

5 Conclusion

Our systematic evaluation for LLM-based GHG emissions extraction suggests that commonly recommended prompt and model choices do not always translate into higher extraction quality. Newer model generations do not consistently improve upon older ones, and the open-weight models we tested trail meaningfully behind the closed-source gpt family, pointing to the value of empirically comparing candidate models on a representative slice of one's own data rather than defaulting to the latest release. A structured prompt following common best practices is outperformed by a simpler question-style prompt, and the effectiveness of tabular context depends on the model-prompt combination. Cutting across these observations, we see a recurring precision-recall trade-off: gains in one metric typically come at the expense of the other, both across model generations and across structured prompt components. gpt-5.2 is a partial exception, achieving higher precision without a proportional loss in recall, but even here the pattern reappears when comparing configurations against one another. Extraction rules are the notable exception on the prompt side, contributing to both precision and recall, whereas role and scope definitions trade one for the other.

From a deployment cost perspective (see Appendix A.10), the best-performing configurations are considerably more expensive than their close alternatives: switching from gpt-4.1 to gpt-5.2 raises cost by roughly 46% for only a small F1 improvement, and the question prompt is 3.8× more expensive than the structured prompt on gpt-5.2 for a similarly modest gain. The prompt-side gap in particular reflects the question prompt's output structure, which restates all 44 scope-year queries and produces roughly 15× more output tokens than the structured prompt's JSON format. Cost differences between design choices can thus dwarf the corresponding performance gains, making the cheaper option often the more rational one.

Limitations

Several caveats should be noted. Our evaluation relies on a small benchmark that is deliberately curated for annotation quality and consists solely of English-language reports, which may constrain how far the findings extend to larger or multilingual corpora. That said, the benchmark covers a broad diversity of document structures and report lengths, mirroring the variability observed in real-world corporate reporting. We believe that the resulting findings should therefore remain meaningful within this scope. Extending our evaluation methodology to additional sustainability corpora is a natural direction for future work, which the released `climatextract` package is designed to facilitate.

The original PDF reports cannot be redistributed through our repository. Readers wishing to reproduce our experiments must retrieve the source documents themselves through the public links accompanying the benchmark dataset (Beck et al., 2025).

More broadly, the design space for LLM-based information extraction, spanning prompt structure, contextual input, and model choice, is vast, and our experiments necessarily cover only a principled subset of it. Findings might therefore not transfer to configurations we did not test.

Our model selection is likewise deliberately narrow: we focus on the closed-source gpt family given its industry prevalence, complemented by gpt-oss and Llama-4 as established representatives of the open-weight landscape. Alternative model families may behave differently.

Use of AI Assistants

The authors used large language model assistants (Claude) to support the writing and revision process. Specifically, AI assistants were used to suggest alternative phrasings, improve sentence-level clarity and explore structural options for the presentation of results. All conceptual contributions, experimental design, data analysis, and final wording decisions were made by the authors. The authors reviewed and verified all AI-assisted output before incorporation into the paper.

References

Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. 2025. SustainableQA: A Comprehensive Question Answering Dataset for Corporate Sustainability and EU Taxonomy Reporting. *ArXiv*, abs/2508.03000.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. What's in a prompt?: A large-scale experiment to assess the impact of prompt design on the compliance and accuracy of llm-generated text annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:122–145.

Jacob Beck, Anna Steinberg, Andreas Dimmelmeier, Laia Domenech Burin, Emily Kormanyos, Maurice Fehr, and Malte Schierholz. 2025. Addressing data gaps in sustainability reporting: A benchmark dataset for greenhouse gas emission extraction. *Scientific Data*, 12(1):1497.

Andreas Dimmelmeier, Hendrik Doll, Malte Schierholz, Emily Kormanyos, Maurice Fehr, Bolei Ma, Jacob Beck, Alexander Fraser, and Frauke Kreuter. 2024. Informing climate risk analysis using textual information - a research agenda. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 12–26, Bangkok, Thailand. Association for Computational Linguistics.

Kerstin Forster, Lucas Keil, Victor Wagner, Maximilian A. Müller, Thorsten Sellhorn, and Stefan Feuerriegel. 2025. Assessing corporate sustainability with large language models: Evidence from europe. Working Paper 202, TRR 266 Accounting for Transparency Working Paper Series.

R. Greene, T. Sanders, L. Weng, and A. Neelakantan. 2022. New and improved embedding model. Accessed: 2026-01-22.

Biyang Guo, He Wang, Wenyilin Xiao, Hong Chen, ZhuXin Lee, Songqiao Han, and Hailiang Huang. 2024. Sample design engineering: An empirical study on designing better fine-tuning samples for information extraction with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 573–594, Miami, Florida, US. Association for Computational Linguistics.

Alice Kwak, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2024. Classify first, and then extract: Prompt chaining technique for information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 303–317, Miami, FL, USA. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Rafik Mankour, Yassine Chafai, Hamada Saleh, Ghassen Ben Hassine, Thibaud Barreau, and Peter Tankov. 2025. Climate finance bench. *Preprint*, arXiv:2505.22752.

Meta AI. 2025. LLaMA 4 Model Card. Accessed: 2026-02-09.

Network for Greening the Financial System. 2024. Information note: Improving greenhouse gas emissions data. Accessed: 2023-10-12.

Julian Neuberger, Lars Ackermann, Han van der Aa, and Stefan Jablonski. 2025. A universal prompting strategy for extracting process model information from natural language text using large language models. In *Proceedings of the 43rd International Conference on Conceptual Modeling, ER 2024*, pages 38–55, Cham. Springer Nature Switzerland.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025b. Introducing gpt-4.1 model family. Accessed: 2026-01-22.

OpenAI. 2025c. Update to gpt-5 system card: Gpt-5.2. Accessed: 2026-01-22.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

WRI and WBCSD. 2011. The greenhouse gas protocol: A corporate reporting and accounting standard - revised version. Accessed: 2026-01-15.

Chongjian Yue, Xinrun Xu, Xiaojun Ma, Lun Du, Zhiming Ding, Shi Han, Dongmei Zhang, and Qi Zhang. 2025. Extract information from hybrid long documents leveraging llms: A framework and dataset. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

A Appendix

A.1 Retrieval experiments

Similarity k	Precision	Recall	F1 Score	True Positives	Cost (€)	Retrieval Failures
3	0.431	0.530	0.475	262	0.274	14
4	0.432	0.568	0.491	284	0.363	10
5	0.432	0.581	0.495	287	0.451	6
6	0.424	0.583	0.491	299	0.541	2
7	**0.431**	**0.608**	**0.502**	**310**	**0.631**	**2**
8	0.433	0.612	0.507	318	0.719	2
9	0.407	0.601	0.485	300	0.807	2

Table A1: Retrival experiments, varying the parameter k, the retrieved number of pages most similar to the query.

We evaluated 12+ retrieval configurations on all 139 reports, varying the number of retrieved pages k and the context window size in Table A1. Using gpt-4o-mini for consistent LLM evaluation, we compared configurations on precision, recall, F1, retrieval failures, and cost. Table A1 reports results for varying k with a context window of 0 adjacent pages, the best-performing window size. Setting $k = 7$ retrieved all ground-truth pages for 137 of 139 reports at competitive precision and F1, minimising retrieval failures. While $k = 6$ performed comparably in our dataset, we selected $k = 7$ as a more robust choice for future datasets.

A.2 Prompt designs

Figure A1 shows the question prompt template. The red dashed box was added when we adopted gpt-5.2, whose output includes Markdown symbols, not parsable by the regex schema. The instruction to reprint the first three questions in the response format was necessary to counteract the model's observed tendency to "forget" early questions in the output.

```
Extract key pieces of information from this sustainability report.
IMPORTANT: Do NOT use any markdown formatting like **bold**, *italic*, or headers.
Write CO2 (not CO₂) when specifying units.
If a particular piece of information is not present, output 'Not specified'.
If the report does not mention if Scope 2 is location-based or market-based,
assume it is Scope 2 location-based. Do not extract it as Scope 2 market-based.

Use the following format:
0. What is the title
1. What are the Scope 1 emissions in 2013
2. What are the Scope 1 emissions in 2014
3. What are the Scope 1 emissions in 2015
4. What are the Scope 1 emissions in 2016
5. What are the Scope 1 emissions in 2017
6. What are the Scope 1 emissions in 2018
7. What are the Scope 1 emissions in 2019
8. What are the Scope 1 emissions in 2020
9. What are the Scope 1 emissions in 2021
10. What are the Scope 1 emissions in 2022
11. What are the Scope 1 emissions in 2023
12. What are the Scope 2 (market-based) emissions in 2013
13. What are the Scope 2 (market-based) emissions in 2014
14. What are the Scope 2 (market-based) emissions in 2015
15. What are the Scope 2 (market-based) emissions in 2016
16. What are the Scope 2 (market-based) emissions in 2017
17. What are the Scope 2 (market-based) emissions in 2018
18. What are the Scope 2 (market-based) emissions in 2019
19. What are the Scope 2 (market-based) emissions in 2020
20. What are the Scope 2 (market-based) emissions in 2021
21. What are the Scope 2 (market-based) emissions in 2022
22. What are the Scope 2 (market-based) emissions in 2023
23. What are the Scope 2 (location-based) emissions in 2013
24. What are the Scope 2 (location-based) emissions in 2014
25. What are the Scope 2 (location-based) emissions in 2015
26. What are the Scope 2 (location-based) emissions in 2016
27. What are the Scope 2 (location-based) emissions in 2017
28. What are the Scope 2 (location-based) emissions in 2018
29. What are the Scope 2 (location-based) emissions in 2019
30. What are the Scope 2 (location-based) emissions in 2020
31. What are the Scope 2 (location-based) emissions in 2021
32. What are the Scope 2 (location-based) emissions in 2022
33. What are the Scope 2 (location-based) emissions in 2023
34. What are the Scope 3 emissions in 2013
35. What are the Scope 3 emissions in 2014
36. What are the Scope 3 emissions in 2015
37. What are the Scope 3 emissions in 2016
38. What are the Scope 3 emissions in 2017
39. What are the Scope 3 emissions in 2018
40. What are the Scope 3 emissions in 2019
41. What are the Scope 3 emissions in 2020
42. What are the Scope 3 emissions in 2021
43. What are the Scope 3 emissions in 2022
44. What are the Scope 3 emissions in 2023

Document:
{context_str}

0. What is the title: Our responsibility. Report {self.min_year}
1. What are the Scope 1 emissions in 2013: <value> <unit>
2. What are the Scope 1 emissions in 2014: <value> <unit>
```

Figure A1: Question prompt template consisting of a question for every scope-year combination in the respective year range 2013-2023 for every scope type Scope 1, 2 (market-based), 2 (location-based) and 3. The text in the red box had to be added to run our pipeline with gpt-5.2.

Figure A2 shows the structured prompt template. Beyond the relevance criteria listed in the main text, the prompt details further extraction rules addressing not-intended transformations (e.g. rejecting percentage increases or intensity ratios).

The output schema in Figure A3 is defined as a Pydantic model in Python and serialised to JSON, specifying a list of KPI entries, each representing one scope-year combination, with the four fields year, kpi_name (the scope), value, and unit. Each field carries a natural-language description and example value that are passed to the LLM as part of the structured prompt to guide extraction; these are omitted from Figure A3 for brevity but are visible in the released code.

A.3 Pipeline Step 4: Parsing

Unit normalization. Companies use different unit variants for the same quantity (e.g. "1,000 t CO2" and "metric kilotons CO2e"). We collect all distinct variants in the dataset and normalise them via a dictionary to one of "t CO2e", "kt CO2e", "Mt CO2e", "MM Mt CO2e", "kg CO2e" or "Other" if unknown. A multiplier per known unit allows standardising all values to "t CO2e" for downstream comparability.

Duplicate resolution. To retain only one entry per scope-year combination, we apply three prioritisation rules sequentially: (1) retain only one entry among duplicates sharing the same value and unit on the same page; (2) among entries with different units, retain the one in "t CO2e"; (3) among entries with matching units, retain the one from the majority page, breaking ties by the earliest page.

A.4 Statistical Testing

To assess the significance of the differences reported in Section 4, we fit fixed-effects regressions on report-level F1 scores. Report-level averaging is used because extractions within the same report cannot be assumed to be independent. Each comparison of interest uses one experimental condition as the reference and enters the other as a dummy variable; the coefficient on the dummy is the estimated F1 difference. We cluster standard errors at both the report and run level: reports differ in complexity, and even at temperature = 0 LLM APIs can produce small run-to-run variation due to backend nondeterminism.

Each test corresponds to a directional hypothesis of the form $F1_A > F1_B$ derived from best-practice recommendations. We compute one-sided p-values against the null that the opposite direction holds. Table A2 reports the coefficient, clustered standard error, and outcome for each comparison.

Two observations are worth flagging. First, the prompt hypothesis $F1_{\text{structured}} > F1_{\text{question}}$ was significantly rejected in the tested direction, meaning the question prompt outperforms the structured prompt at $p < 0.05$. Second, comparisons of gpt-5.2 against gpt-4o and gpt-4.1 do not reach significance, so although gpt-5.2 achieves the highest F1 in absolute terms, the gap to the closest earlier generations is not statistically supported.

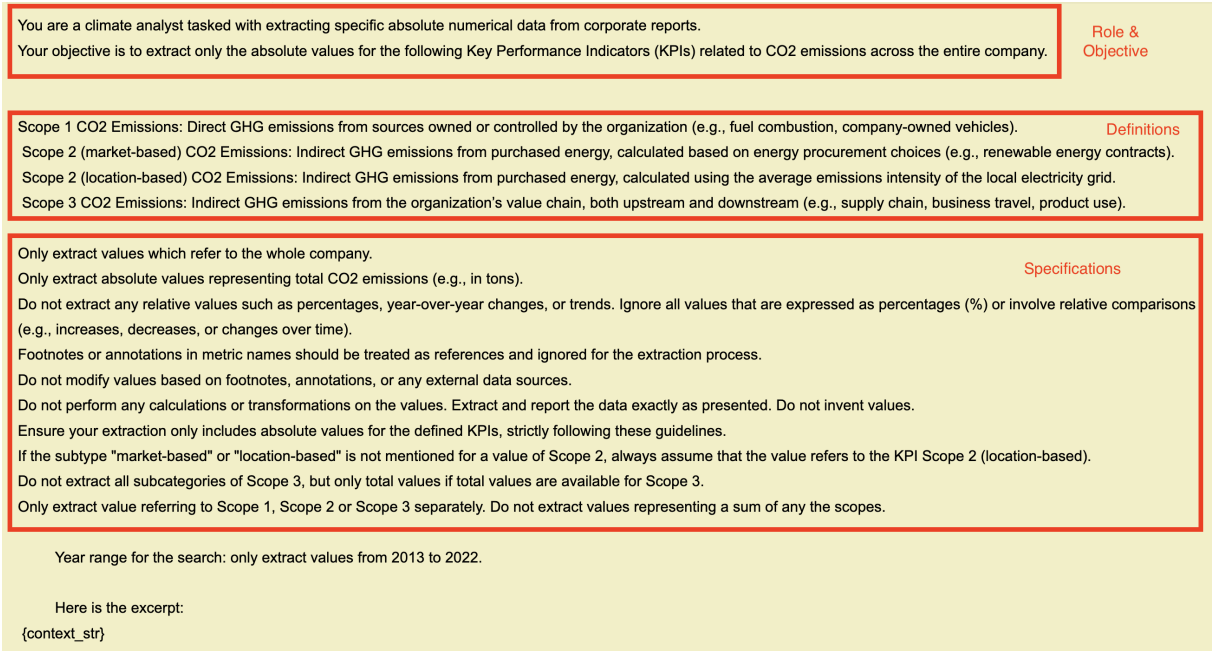


Figure A2: Structured prompt template consisting of role and objective, definitions of scope types, extraction rules and desired year range.

```
{
  "$defs": {
    "KpiEntry": {
      "type": "object",
      "properties": {
        "year": {"type": ["integer", "null"]},
        "kpi_name": {"type": ["string", "null"],
                     "enum": ["1", "2 (market-based)
                         ", "2 (location-based)",
                         "3", null]},
        "value": {"type": ["number", "null"], "
            minimum": 0},
        "unit": {"type": ["string", "null"]}
      }
    }
  },
  "type": "object",
  "properties": {
    "KPI_Entries": {"type": "array", "items": {"$ref
        ": "#/$defs/KpiEntry"}}
  },
  "required": ["KPI_Entries"]
}
```

Figure A3: Output schema serialised from Pydantic model

Comparison	Directional hypothesis	Coefficient (SE)	Outcome
Prompt	$F1_{\text{structured}} > F1_{\text{question}}$	-0.0085^{**} (0.0031)	not supported
Context	$F1_{\text{text+table}} > F1_{\text{text}}$	0.0167^{***} (0.0037)	supported
Model	$F1_{\text{gpt-5.2}} > F1_{\text{gpt-4o}}$	0.0064 (0.0045)	not supported
Model	$F1_{\text{gpt-5.2}} > F1_{\text{gpt-4.1}}$	0.0048 (0.0046)	not supported
Model	$F1_{\text{gpt-5.2}} > F1_{\text{gpt-5}}$	0.0077^{**} (0.0041)	supported
Model	$F1_{\text{gpt-5.2}} > F1_{\text{Llama-4}}$	0.1485^{***} (0.0227)	supported
Model	$F1_{\text{gpt-5.2}} > F1_{\text{gpt-oss}}$	0.2713^{***} (0.0246)	supported

Table A2: Directional hypothesis tests on report-level F1 scores from a fixed-effects regression. Each row reports the coefficient on the dummy variable corresponding to the first condition in the directional hypothesis, with clustered standard errors (report and run) in parentheses. Significance: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

A.5 Output Format for gpt-oss-120b

Section 4.1 mentions that 85.8% of gpt-oss-120b's errors on non-NA reports stem from output-formatting issues our regex parser cannot handle, rather than incorrect values. To contextualize this, we compare against Llama-4 (alternative open-weight model) and gpt-4.1 (second-best model): only 13.1% and 3.8% of their errors are formatting-driven. This contrast underscores that gpt-oss-120b's apparent underperformance is largely an artifact of its output format. We therefore re-run gpt-oss-120b under the structured JSON prompt to assess its capability independent of output formatting; Table A3 reports the results, averaged over 5 runs.

Prompt Type	Precision	Recall	F1
question	0.812 (0.040)	0.281 (0.042)	0.416 (0.047)
structured	0.760 (0.013)	0.814 (0.017)	0.786 (0.012)

Table A3: gpt-oss-120b performance by output format, averaged over 5 runs. Standard deviations in parentheses.

A.6 Ablation of output format

We perform an ablation study with respect to the output format by testing the question prompt with the same JSON output format as the structured prompt. The results are given in Table A4. The JSON output format leads to a decrease in F1 score from 0.823 to 0.794. This indicates that the output format does not explain the performance gap between the question and structured prompts and, if anything, slightly hurts the question prompt's performance.

Prompt components	Precision	Recall	F1 Score
question prompt	0.779 (0.015)	0.878 (0.084)	0.823 (0.039)
question prompt with JSON output	0.777 (0.031)	0.812 (0.019)	0.794 (0.022)
structured prompt	0.760 (0.008)	0.860 (0.011)	0.807 (0.006)

Table A4: Averages and standard deviations of 5 runs of ablations of prompt output format. The question prompt with JSON output format performs worse than the question prompt with regex and the structured prompt.

A.7 Report-level analysis

We categorise each report's extractions under both prompts and track how individual reports shift between categories. Reports are first split by whether the ground-truth benchmark contains disclosed emissions (non-NA reports) or not (NA reports). NA reports are classified as either correct (no values returned) or incorrect (at least one irrelevant value extracted). Non-NA reports are classified as completely correct, partially correct, or completely incorrect, with a separate retrieval-failure category for reports where the retrieved pages did not contain the relevant ground-truth information. Figure A4 shows the resulting transitions. Zooming in on the shifts: 4 NA reports switch from incorrect under the question prompt to correct under the structured prompt, but 2 correctly-handled NA reports shift the other way, extracting irrelevant values such as Scope 3 subcategory totals. Among non-NA reports, 2 of 139 are classified as retrieval failures where ground-truth values were spread across pages and at least one relevant page was not retrieved. The persistence of subcategory-

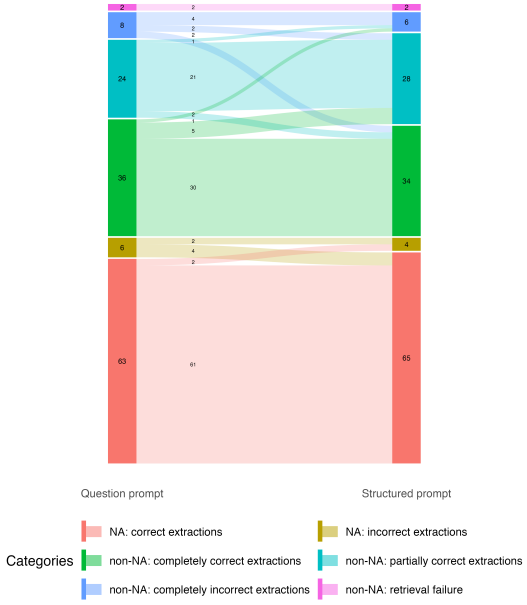


Figure A4: Diagram comparing report-level extraction results between the question prompt and the structured prompt when using gpt-5.2 model. Most reports remain in the same category. Improvements for some are offset by regression for other reports.

extraction errors under the structured prompt indicates that the LLM does not consistently follow the extraction rules included in the prompt, unlike the more successful adherence reported by Mankour et al. (2025); Neuberger et al. (2025). Another recurring error is the model's tendency to interpret footnote numerals as part of a value.

A.8 Second-best model

Prompt Type	Input Mode	Precision	Recall	F1 Score
question	text	**0.674** (0.040) [0.642, 0.750]	**0.953** (0.027) [0.916, 1.000]	**0.789** (0.034) [0.764, 0.857]
question	text+table	0.664 (0.095) [0.504, 0.750]	0.922 (0.168) [0.451, 1.000]	0.769 (0.120) [0.476, 0.857]
structured	text	0.582 (0.031) [0.541, 0.624]	0.853 (0.014) [0.838, 0.869]	0.692 (0.021) [0.667, 0.725]
structured	text+table	0.587 (0.020) [0.554, 0.602]	0.836 (0.053) [0.744, 0.872]	0.689 (0.031) [0.635, 0.713]

Table A5: Averaged results of 5 runs of prompt and context experiments with gpt-4.1, accompanied by standard deviation and range. A simple question-style prompt outperforms the structured prompt. Adding tabular context deteriorates results for both prompt types.

Table A5 reports the prompt-context experiments repeated with the second-best model, gpt-4.1. In contrast to gpt-5.2, adding tabular context weakens performance for both prompt types, indicating that the benefit of tabular context is not robust across models.

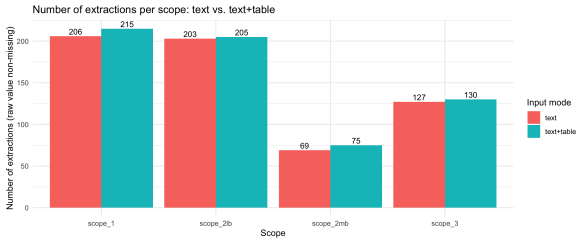


Figure A5: Number of non-missing extractions per emissions scope, comparing text-only and text+table context for gpt-5.2 model and question prompt type. Text+table yields equal or higher counts across all scopes, with the largest relative gains for Scope 2 market-based and Scope 3.

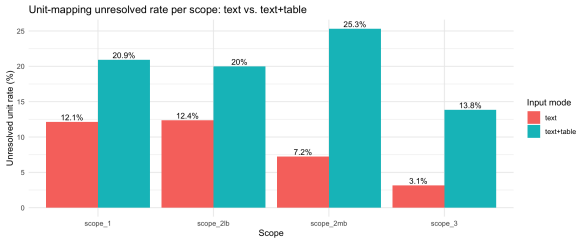


Figure A6: Rate of extracted units unresolved to a known category (mapped to *Other*) per scope, text vs. text+table context for gpt-5.2 model and question prompt type. Text+table increases the unresolved rate across every scope.

unit_cat	text (%)	text+table (%)
t CO2e	68.5	59.2
kt CO2e	13.2	12.3
Other	9.8	19.7
Mt CO2e	7.5	8.0
MM Mt CO2e	1.0	0.0
kg CO2e	0.0	0.8

Table A6: Distribution of resolved unit categories under text and text+table context for gpt-5.2 model and question prompt type. Text+table shifts mass from *t CO2e* toward *Other*, consistent with the model dropping the "e" suffix when the unit-less structured table is included.

A.9 Tabular context: recall gains and unit-resolution

This appendix provides scope-level detail supporting the effects of tabular context (text+table) on extraction recall and unit resolution discussed in Section 4.3.

Table A6 reports the full distribution of resolved unit categories under text-only and text+table context, including the share falling into *Other* when no key in the unit mapping matches. Figure A6 isolates this unresolved share and breaks it down by scope. Scope 1 and Scope 2lb track each other closely in both settings because they are conventionally reported as adjacent line items sharing a single unit descriptor (e.g., one column header spanning both rows in a summary table); a mapping miss for that shared string therefore fails both scopes simultaneously. Case numbers are too small for a detailed analysis of Scope 2mb: Almost all unresolved units from Scope 2mb under text+table originate from a single PDF file. Scope 3 remains lowest throughout, consistent with its typically being reported in a separate category-breakdown table rather than alongside Scope 1/2, and so not exposed to the same shared-unit failure mode.

A.10 Cost-benefit analysis

Model	F1 Score	Costs (€)	Prompt Tokens	Output Tokens
gpt-4.1	0.775	9.51	1469340	821556.0
gpt-5.2	0.823	13.92	1468518	810563.8

Table A7: Cost-benefit comparison between two best models, gpt-4.1 and gpt-5.2, using the question prompt and text-only context. Spending 46% more on average leads to a 6% boost in performance. Note that the figures presented refer to averages of 5 runs of 139 reports for a given setting.

Prompt Type	Input Mode	F1 Score	Costs (€)	Prompt Tokens	Output Tokens
structured	text	0.807	3.59	1622519	53699.2
structured	text+table	0.787	3.72	1709280	51940.0
question	text	0.823	13.92	1468518	810563.8
question	text+table	0.835	14.70	1684756	839233.4

Table A8: Cost-benefit analysis for gpt-5.2 model for different prompt and context type settings. Using the question-style prompt instead of the structured prompt for text-only context requires 36% more tokens overall, driven by approximately 15 times more output tokens, while yielding a 2% higher F1 score. Note that the figures presented refer to averages of 5 runs of 139 reports for a given setting.

Tables A7 and A8 report the cost figures underlying the discussion in the conclusion. Figures represent averages of 5 runs over 139 reports; embedding costs are excluded, since embedding is a one-time cost per report.

Table A7 compares the two best models under the question prompt and text-only context: switching from gpt-4.1 to gpt-5.2 increases cost by 46% on average for a 6% F1 improvement (0.775 → 0.823).

Table A8 compares prompt-context combinations for gpt-5.2. Adding tabular context adds only 0.78 EUR to the question prompt and 0.12 EUR to the structured prompt, but its effect on F1 is direc-

tional: $+1.4\%$ for the question prompt and -2.4% for the structured prompt.

Who Warmed the Archives? LLMs Overestimate Historical Warmth

Claudiu Creanga[2,3]*, **Liviu P. Dinu**[1,3]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
University of Bucharest, Romania
claudiu.creanga@s.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Historical archives are an under-used source for extending the instrumental climate record backward in time, and LLMs offer a way to extract the indices climatologists derive by hand. Beyond measuring how well systems extract this signal, we check whether their errors are safe to use for cross-century comparison, since a good correlation score does not rule out systematic, era-linked bias. Comparing lexical baselines, fine-tuned historical transformers, and LLM prompting on the Pfister temperature index across five centuries of German text, lexical methods beat every fine-tuned transformer we test, including one pretrained from scratch on historical German ($r = -0.016$). All six LLMs we test (Gemini 2.5 Flash, GPT-5-mini, DeepSeek v4 Flash, Claude Sonnet 4.6, Qwen3.7-Plus, Kimi-K2.6-Fast) show a warm bias that grows with calendar year, with the same sign in every model (slopes $+0.13$ to $+0.34$/century, $p < 0.01$). The effect is modest in size ($r^2 \approx 0.01$–0.05) but consistent across six independently developed models. The best-correlated of the six, Gemini 2.5 Flash, matches the best lexical correlation ($r = 0.32$) at double the error. An ablation stripping explicit dates and calendar-era markers from the quotes leaves this trend essentially unchanged, favoring an anachronistic present-day prior over the model correctly inferring the quote's era. Correlation alone is thus insufficient for vetting an LLM as a historical-climate-index oracle.

> But where are the snows of yesteryear?
>
> François Villon

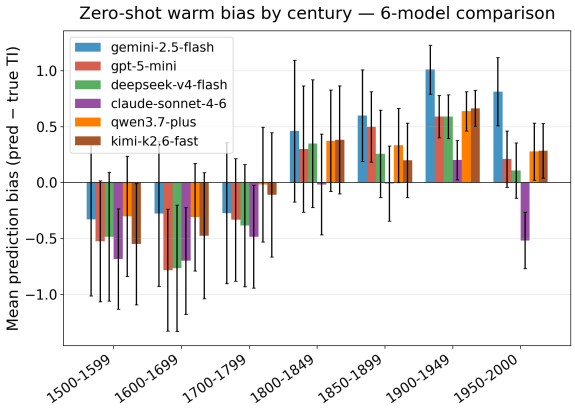


Figure 1: Mean zero-shot prediction bias by century, all six audited models, identical 702-row sample: a cold bias before 1800 flips to a warm bias after, in every model.

1 Introduction

Historical documentary archives (chronicles, weather diaries, and newspaper reports spanning five centuries) record climate information via the semi-quantitative ordinal indices long used in historical climatology (Pfister, 1984; Glaser and Riemann, 2009; Glaser and Kahle, 2020). NLP, and increasingly LLM prompting, is an obvious way to scale up this extraction. Most evaluations of this kind stop at whether a system can extract usable climate signal from old text at all, typically measured with a correlation or F1 score. We additionally check whether that signal can be trusted for the **cross-century comparisons** a paleoclimatologist would actually want to make with it.

We compare lexical baselines, historical-domain fine-tuned transformers (Europeana BERT, GHisBERT, Historic Multilingual BERT), and zero-/few-shot LLM prompting (Gemini 2.5 Flash (Gemini Team, 2025), GPT-5-mini (OpenAI, 2026), DeepSeek v4 Flash (DeepSeek-AI, 2026), Claude Sonnet 4.6 (Anthropic, 2026), Qwen3.7-Plus (Qwen Team, 2026), Kimi-K2.6-Fast (Kimi Team, 2026)) on predicting the Pfister temperature index from German historical text (1500–2022), then audit whether the best LLM's errors are random or systematically structured by era across all

*Corresponding author.

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 41–50
October 28, 2026 ©2026 Association for Computational Linguistics

six models. GHisBERT, pretrained on historical German from scratch, scores worse than random ($r = -0.016$) against a plain TF-IDF baseline ($r = 0.32$). All six LLMs show a warm bias that grows significantly with calendar year. Gemini, the best-correlated of the six, matches that lexical correlation while carrying about twice the error (Figure 1). Exact figures appear in §5.

These results motivate three questions: where climate signal lives in historical text (**RQ1**), whether LLM prompting can match specialized fine-tuning (**RQ2**), and whether an LLM's miscalibration here reflects a systematic, temporally structured anachronistic prior, or is better explained as unstructured noise (**RQ3**). They yield five contributions:

- A controlled comparison of lexical, fine-tuned-transformer, and LLM-prompting approaches on five centuries of German/Latin text, showing lexical methods beat every fine-tuned transformer we test, including one purpose-built for this exact language variety.
- Evidence that correlation alone can be a misleading fitness measure for an LLM's climate-index accuracy: it can tie a lexical baseline's r while its absolute error runs twice as large, because its errors follow a systematic, directional pattern.
- A cross-model audit showing this systematic error is temporally structured and replicates across **six independently developed model families** (Google, OpenAI, DeepSeek, Anthropic, Alibaba, and Moonshot AI), suggesting it is a general property of current LLMs, not an idiosyncrasy of a single model.
- An **era-cue-stripping ablation** showing this bias survives, essentially unchanged, once explicit dates and calendar-era markers are removed from the prompt. This points to an anachronistic prior: the model does not appear to read the quote's stated date correctly and adjust its prediction accordingly.
- A secondary case study on structured event extraction from multilingual medieval records, showing **Latin** quotes are not easier for the model than dialectal **German** despite Latin's greater standardization.

2 Related Work

A growing line of work asks whether contemporary LLMs can be trusted to represent, or reason about, periods predating their training distribution's center of mass. Underwood et al. (2025) show that prompting or light fine-tuning does not convincingly reproduce period prose style, arguing that simulating a historical perspective may require pretraining on period corpora. Cuscito et al. (2024) and Cuscito et al. (2025) similarly find a model pretrained on historical text (MacBERTh) beats general-purpose BERT and ChatGPT on Early Modern English fill-in-the-blank tasks, while the reverse holds on contemporary items, showing that prompting alone does not confer period competence. Levchenko (2025) document a "modernization trap" and paradoxical "over-historicization" in multimodal-LLM OCR of 18th-century Russian books, arguing models treat "old" as a single undifferentiated space instead of a specific period. Hutchinson (2024) frames such errors as historiographical, not merely as simple bugs, a challenge that echoes recent findings on how NLP systems struggle to reconcile divergent historical narratives (Creanga et al., 2026). We extend this evidence to a new modality (climate-index prediction) and setting (historical German text). Unlike Levchenko's coarse old-vs.-new binary, our bias increases gradually across century-scale buckets, consistent with a graded sense of period.

A related literature studies what happens when a model's provided context and pretraining-derived ("parametric") beliefs disagree. Longpre et al. (2021) formalize this as a *knowledge conflict*, showing QA models over-rely on parametric knowledge even against contradicting context. Neeman et al. (2023) train models to disentangle context-grounded from parametric answers, and Xu et al. (2024) survey the resulting taxonomy. Li et al. (2025) find the same attention heads can promote both signals simultaneously, and Yamin et al. (2025) show LLMs default almost exclusively to parametric knowledge under explicit counterfactual premises. We adopt this framing here: a quote's date and content make up the context. When the model over-predicts pre-industrial warmth for later (but still pre-1900) quotes, that pattern looks like a parametric prior that "the world has warmed" overriding the context.

A third literature studies LLMs' general-purpose sense of time. Dhingra et al. (2022) show LMs

trained on a fixed snapshot are miscalibrated on facts that change over time. Zhao et al. (2024) find LMs often default to a year *earlier* than their true training cutoff ("temporal chaos"), with alignment able to shift this anchor either direction. Piryani et al. (2025) survey temporal QA more broadly. Our audit is a domain-specific instance of this problem, though our target differs from recall of time-sensitive facts: **we test whether a model's numeric climate judgment is anchored to the source quote's true century, or defaults toward a present-day prior**.

Our fine-tuned baselines draw on BERT-style models pretrained on historical text. Schweter et al. (2022) (dbmdz) release hmBERT, a multilingual BERT pretrained on OCR'd Europeana/British Library text for German, English, French, Swedish, and Finnish. We also use its earlier German-only sibling trained on noisier Europeana data (`dbmdz/bert-base-german-europeana-cased`).[1] Beck and Köllner (2023) instead pretrain GHisBERT from scratch on historical German back to 750 CE, arguing that continuing pretraining from a modern German checkpoint scales poorly to text this old. We include it as the baseline pretrained across our corpus's full historical depth.

The indices we predict come from historical climatology, a field that reconstructs pre-instrumental climate from documentary evidence such as chronicles and diaries. Pfister (1984) introduces the semi-quantitative ordinal-index approach that remains dominant in the field, converting qualitative chronicle/diary/record descriptions into a seven-point per-month scale calibrated against instrumental overlap. Glaser and Riemann (2009) apply this to produce the thousand-year German/Central European temperature reconstruction underlying our TI task, and Glaser and Kahle (2020) apply the same methodology to a Historical Humidity Index since 1500, underlying our HHI task. Both, with the underlying quotations, are hosted in tambora.org (Riemann et al., 2015). Unlike this prior work, our contribution is testing whether NLP models can recover these expert-assigned indices directly from the quotations.

Finally, our audit is motivated by the growing use of LLMs within climate research itself. Nabavi et al. (2026) survey such applications, highlighting both the opportunities and the social, environmental, and epistemic risks they introduce. Our audit contributes one concrete instance of such a risk, systematic misjudgment of historical climate conditions, for a plausible near-term application (LLM-assisted digitization) of exactly the models we test.

3 Data

All experiments use quotations from **historical German** (with occasional French and Latin) chronicles, newspapers, and administrative weather records, together with their associated documentary climate indices. The quotations and indices originate from the *tambora.org* collaborative research environment for climate and environmental history (Riemann et al., 2015), a database of dated, geo-referenced climate-relevant text passages linked to derived index values. The two index families we predict (a monthly Pfister-style temperature index and a monthly historical humidity index) are constructed using the semi-quantitative ordinal-index methodology of Pfister (1984) (§2), as applied to Germany and Central Europe by Glaser and Riemann (2009) (temperature) and Glaser and Kahle (2020) (drought/humidity).

Our primary corpus pairs each quotation with its monthly Pfister temperature index (TI), an ordinal anomaly score $\mathrm{TI} \in \{-3, \ldots, +3\}$ relative to a local historical baseline, which we treat as a continuous regression target under mean-squared-error loss. We use a chronological split, training on older text and evaluating on more recent text, to mimic the realistic deployment scenario of extrapolating index values into eras with progressively different source material. Table 1 gives split sizes and year ranges. Text is predominantly German, with historical spelling variation typical of the period (e.g. *schney*/*schnee*, *fluth*/*flut*). A small number of later (19th–20th century) newspaper excerpts are in French, and the earliest quotations occasionally contain Latin. We do not filter these out, so all reported numbers reflect this realistic multilingual mix. We additionally construct an expanded training variant that adds 45 further pre-1850 rows (+1%) drawn from a larger raw extract that also contains pre-1500 quotations without a monthly TI label (see the medieval corpus described below). We report this 800-dataset variant alongside the original where relevant, since the additional rows do not meaningfully change downstream results (§5).

[1] Released without an accompanying paper. See `https://github.com/dbmdz/berts`.

Split	Records	Years
Train	4,022	1500–1849
Validation	558	1850–1899
Test (TI)	502	1900–2022
Test (HHI)	666	1900–2026

Table 1: Main corpus splits, temperature index (TI) and humidity index (HHI). Train/validation are shared. Only the test period and size differ.

Portion	n	Years
Pre-1900 (stratified, fresh)	300	1500–1899
1900–2000 (reused test set)	402	1900–2000
Total	**702**	1500–2000

Table 2: Century-stratified bias-audit sample, used identically across all six audited LLMs.

To test this across the full historical range, we construct a second, purpose-built sample spanning 1500–2000 (Table 2). For the pre-1900 portion, we draw a fresh stratified sample of $n = 20$ quotations per gold-TI sign (cold: $\mathrm{TI} < 0$; neutral: $\mathrm{TI} = 0$; warm: $\mathrm{TI} > 0$) within each of five century-scale buckets (1500s–1850s), for 300 rows total, so that any cross-century difference in model bias cannot simply be an artifact of the gold-label distribution shifting across eras. For the 1900–2000 portion, we reuse the 402 held-out zero-shot test predictions from the main corpus instead of resampling, for cost efficiency when replicating the audit across six LLM providers (§4). This reused portion is *not* sign-balanced. We treat this pre-/post-1900 sampling asymmetry as a limitation (§6) and control for it throughout by reporting bias separately within each true-TI-sign subset, so that comparisons across centuries are always like-for-like on label composition.

As a secondary, generalization check on a second climate variable, we pair the same quotations with the monthly Historical Humidity Index ($\mathrm{HHI} \in [-4, +4]$), which combines a historical drought index and a historical wet index derived by the same ordinal-index methodology (Glaser and Kahle, 2020). The chronological split (Table 1, HHI row) mirrors the main corpus but extends the test period to 2026 to match the available HHI series.

Because pre-1500 quotations in the raw *tambora.org* extract lack a monthly TI label, we build a separate, seasonally resolved variant for the **medieval period** (1000–1499): each quotation is paired with a seasonal (winter/spring/summer/autumn) Pfister TI instead of a monthly one, split chronologically into 477 train (1000–1349), 166 validation (1350–1424), and 219 test (1425–1499) season-years. This subset is small, and the text is sparse (medieval Latin and Old/Middle/Early New High German, often with only 1–3 short quotations per season), so we read any result on this split as suggestive at best. Season labels are themselves imbalanced (mean TI: summer $+0.54$, autumn -0.26, spring -0.39, winter -0.45), which looks like a genuine seasonal-cycle signal in the gold data and not a sampling artifact.

As a **secondary case study** extending our signal-location question from index regression to structured information extraction, we also use a corpus of 485 unique medieval quotations (c. 800–1532; 104 Latin, 381 German or other) manually annotated with 2,769 discrete climate/hydrology/society events. Each event follows a five-field schema (category, node label, value label, location, year). The category, node label, and value label fields are each constrained to a controlled codebook of 6 categories, 68 node labels, and 33 value labels. We hold out 60 quotations (seed 42) as a fixed few-shot example pool and evaluate on the remaining 425.

4 Experimental Setup

For the main temperature-index and drought/humidity-index regression tasks, we compare four modeling approaches: lexical baselines, fine-tuned historical transformers, zero-shot LLM prompting, and few-shot LLM prompting. As lexical baselines, we fit a word-level TF-IDF vectorizer (unigrams and bigrams, minimum document frequency 3, maximum document frequency 0.8, sublinear term-frequency scaling) followed by ridge regression, with the regularization strength chosen by five-fold cross-validation over 100 log-spaced values between 10^{-3} and 10^{3}. We also fit a character n-gram TF-IDF vectorizer (word-boundary-aware, n=3–5, minimum document frequency 3, maximum document frequency 0.95) followed by gradient-boosted trees (500 trees, max depth 6, learning rate 0.05, subsample and column-subsample 0.8), both with and without an additional normalized-year feature appended to the character representation. As fine-tuned transformer baselines, we take three BERT-style encoders pretrained on historical or historically

adjacent German text: Europeana BERT, GHisBERT, and Historic Multilingual BERT (§2). We attach a single-output regression head under mean-squared-error loss to each. All three are fine-tuned identically for 3 epochs at learning rate 3e-5, batch size 16, weight decay 0.01, with the best checkpoint selected by validation MAE.

As general-purpose LLM baselines, we prompt six models: Gemini 2.5 Flash, GPT-5-mini, DeepSeek v4 Flash, Claude Sonnet 4.6, Qwen3.7-Plus, and Kimi-K2.6-Fast. All six are queried zero-shot with an identical system prompt asking for a single Pfister-scale integer or decimal in $[-3, +3]$ given the (possibly archaic-spelled) German quote, truncated to 800 characters. The raw response is parsed for the first signed number and clamped to range. All models are queried at `temperature=0` where supported. For reasoning models, chain-of-thought tokens are discarded and only the parsed numeric prediction is kept.

We additionally test **few-shot prompting** for Gemini 2.5 Flash, drawing examples from the pre-1850 training split under two retrieval strategies (uniformly random, and nearest-neighbor by cosine similarity over the same character n-gram TF-IDF representation used for the lexical baselines), at 3-shot random, 5-shot random, and 3-shot similarity, all on the full 502-row test set. The zero-shot condition is evaluated on the 402-row reused subset described in §3.

To probe this, we regress each model's bias (prediction minus true index) against the calendar year of the source quote, over the stratified 702-row sample described in §3, both pooled and within each true-index-sign subset (since the pre- and post-1900 portions differ in sign-balance). We run this identically for all six models, so any difference in trends reflects only which model answered. To adjudicate between the model correctly inferring a quote's era from context and an anachronistic present-day prior applied regardless of context, we additionally construct an era-cue-stripped variant of the same 702-row sample: every explicit four-digit year (1000–2099) is replaced with a fixed placeholder, and explicit Latin/German calendar-era markers (*Anno Domini*, *Ao.*, *n. Chr.*, *im Jahre*, etc.) are removed. We re-run the identical zero-shot prompt on the redacted text for four of the six models (Gemini 2.5 Flash, GPT-5-mini, DeepSeek v4 Flash, and Claude Sonnet 4.6). We scope this ablation to unambiguous numeric/calendar cues, and do not attempt automatic proper-noun redaction via named-entity recognition. This is because we found that a German NER model (spaCy `de_core_news_sm`) mistags common tokens in this archaic orthography badly enough (e.g. tagging *Anno Domini* or the Latin month name *Majo* as `PERSON`) that using it would inject errors in our ablation study.

For the secondary medieval event-extraction case study, we score predicted vs. gold events as multisets (`Counter` intersection) instead of one-to-one bipartite alignment, since events within a quote typically share one implied date. We compute year error once per quote as the absolute difference between median gold year and nearest predicted year. The codebook (the closed category/node-label/value-label vocabulary from §3) is given verbatim in the system prompt but not enforced via schema, so codebook adherence measures genuine vocabulary reuse, since nothing forces compliance. We compare zero-shot against 3-shot random, with examples from a fixed, held-out pool of 60 quotations (seed 42).

Across all regression tasks we report MAE, RMSE, and Pearson correlation on validation and test splits. For the bias audit we report mean bias per century bucket and the slope and significance of bias regressed against year. For event extraction we report strict-match F1, category- and node-label-only F1, year MAE, location hit rate, and codebook adherence.

5 Results

Table 3 compares all main-corpus models on the temperature-index task. Across both the original and expanded training sets, the two TF-IDF lexical baselines outperform all three fine-tuned historical transformers on test-set correlation: the best transformer, Historic Multilingual BERT, reaches $r = 0.2614$, while even the weakest lexical baseline still reaches $r = 0.2771$ (char n-gram XGBoost with a year feature, expanded set), and the strongest, word-level ridge, reaches $r = 0.3225$. The best absolute calibration overall belongs to the character n-gram XGBoost with a year feature on the expanded set (test MAE$= 0.8596$). GHisBERT, the only baseline pretrained from scratch on historical German going back to Old High German, achieves a test correlation of $r = -0.0161$, indistinguishable from noise, despite being the domain-matched model for exactly this language variety. Figure 2 visualizes this ranking directly: GHis-

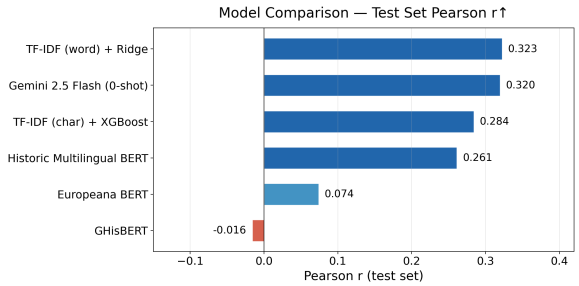


Figure 2: Test-set Pearson r across all main-corpus models.

BERT is the only model below zero, and the lexical baselines sit close to Historic Multilingual BERT well above it.

GHisBERT's test correlation remains near zero despite fine-tuning under the same regime as the other two transformers. Separately, expanding the training set by 45 rows (+1%) via the [+800] variant produces no meaningful change in either the lexical or transformer results (e.g. word ridge test MAE moves from 0.9012 to 0.8905, test r from 0.3225 to 0.3091).

Table 4 turns to whether prompting a general-purpose LLM can substitute for either the lexical baselines or the fine-tuned transformers. Gemini 2.5 Flash's zero-shot correlation on the evaluated 402-row subset ($r = 0.3197$) matches the TF-IDF word ridge baseline ($r = 0.3225$) almost exactly. But its MAE (1.7192) is nearly $2\times$ the ridge baseline's (0.9012), the central dissociation motivating RQ2: a metric that looks tied on correlation conceals a much larger absolute-error gap. The reason is that Gemini's errors form a right-skewed distribution of over-predictions, with mean error $+0.938$ and a tail extending to $+6$ index points on the coldest months, instead of symmetric scatter around the true value. The worst individual errors confirm the pattern: Gemini almost never predicts an index at or below -1 unless the quote contains unambiguous cold vocabulary (*Frost*, *Eis*, *Schnee*). A July 1997 heatwave report and a July 1983 flood report, both true index $+3$, are predicted around -1, while a January 1942 ice-jam report (true index -3) is predicted at $+2.5$.

Few-shot prompting steadily **narrows this gap** without closing it. Moving from zero-shot to 3-shot random, 5-shot random, and 3-shot similarity-retrieved examples lowers MAE from 1.72 to 1.50, 1.46, and finally 1.44, while correlation rises from 0.32 to 0.35, 0.37, and 0.41 respectively. Similarity-based retrieval beats random retrieval at the same shot count ($r = 0.4140$ vs. $r = 0.3537$ at 3-shot), suggesting relevant grounding examples help the model recalibrate its warm bias more than example count alone. The best few-shot condition now exceeds the best fine-tuned transformer's correlation (0.4140 vs. 0.2614) and approaches the TF-IDF ceiling (0.3225). It still trails every lexical and transformer baseline on absolute error, though, since its MAE (1.44) remains well above even the weakest transformer's test MAE (0.90).

Table 5 turns the same audit in the opposite direction: instead of comparing models on a single, mostly-20th-century test period, we regress each model's zero-shot bias against the calendar year of the source quote, across the stratified 702-row sample spanning 1500–2000 (§3), for all six models. Five of the six (Gemini, GPT-5-mini, DeepSeek v4 Flash, Qwen3.7-Plus, and Kimi-K2.6-Fast) drift from a slight cold bias on 16th–18th-century text toward an increasingly warm bias on 19th–20th-century text: Gemini moves from -0.328 (1500s) to $+1.011$ (1900–1949), GPT-5-mini from -0.525 to $+0.591$, DeepSeek from -0.483 to $+0.591$, Qwen3.7-Plus from -0.300 to $+0.638$, and Kimi-K2.6-Fast from -0.550 to $+0.664$, before all five ease somewhat in 1950–2000. Claude Sonnet 4.6 shows a much flatter version of the same early drift, moving from -0.683 in the 1500s to only -0.017 by 1800–1849 and never clearly turning positive in the older centuries. Unlike the other five, it then swings back to a pronounced cold bias in 1950–2000 (-0.517) after a brief positive bump in 1900–1949 ($+0.202$).

A linear regression of bias against year (Table 6) confirms this trend is statistically robust, not an artifact of century-bucket binning. It replicates across all six models, all on the exact same 702 quotes and system prompt ($n = 702$ throughout): Gemini's bias increases by $+0.344$ index points per century ($r = 0.212$, $p < 0.0001$), GPT-5-mini's by $+0.272$ ($p < 0.0001$), DeepSeek's by $+0.250$ ($p < 0.0001$), Kimi-K2.6-Fast's by $+0.257$ ($p < 0.0001$), Qwen3.7-Plus's by $+0.201$ ($p < 0.0001$), and Claude's by $+0.129$ ($p = 0.0054$). **Every slope is positive** and significant, pointing to a shared underlying mechanism across all six model families. Overall mean bias is not uniformly positive, though: pooled across the sample, mean bias ranges from $+0.553$ (Gemini) down to -0.199 for Claude, the only model net cold-biased overall while still warming significantly with year, show-

Model	Val MAE↓	Val RMSE↓	Val r↑	Test MAE↓	Test RMSE↓	Test r↑
TF-IDF (word) + Ridge	0.8329	1.0264	0.3368	0.9012	1.1423	0.3225
TF-IDF (char) + XGBoost	0.8225	1.0179	0.3443	0.8687	1.1245	0.2844
TF-IDF (char) + XGBoost + year	0.8305	1.0260	0.3246	0.8762	1.1239	0.2905
TF-IDF (word) + Ridge [+800]	0.8282	1.0219	0.3384	0.8905	1.1342	0.3091
TF-IDF (char) + XGBoost [+800]	0.8325	1.0354	0.3096	0.8625	1.1145	0.2707
TF-IDF (char) + XGBoost + year [+800]	0.8337	1.0348	0.3134	**0.8596**	**1.1117**	0.2771
Europeana BERT	0.8419	–	0.1895	0.9034	–	0.0741
GHisBERT	0.8467	–	0.0138	0.8841	–	−0.0161
Historic Multilingual BERT	0.8436	–	0.2740	0.8652	–	**0.2614**

Table 3: Temperature-index regression: lexical baselines (original dataset and the [+800]-expanded training variant, char n-gram = 3–5) vs. fine-tuned historical transformers (original dataset only).

Condition	n	MAE↓	RMSE↓	r↑
Zero-shot	402	1.7192	2.0535	0.3197
3-shot random	502	1.4975	1.8407	0.3537
5-shot random	502	1.4574	1.7897	0.3745
3-shot similarity	502	1.4404	1.7619	**0.4140**

Table 4: Gemini 2.5 Flash, zero-shot vs. few-shot, on the temperature-index test set.

Century	n	Gemini	GPT-5	DeepSeek	Claude	Qwen	Kimi
1500–1599	60	−0.328	−0.525	−0.483	−0.683	−0.300	−0.550
1600–1699	60	−0.278	−0.783	−0.767	−0.700	−0.308	−0.475
1700–1799	60	−0.273	−0.333	−0.383	−0.483	−0.017	−0.108
1800–1849	60	+0.462	+0.300	+0.350	−0.017	+0.375	+0.383
1850–1899	60	+0.600	+0.500	+0.258	−0.008	+0.333	+0.200
1900–1949	253	+1.011	+0.591	+0.591	+0.202	+0.638	+0.664
1950–2000	149	+0.813	+0.211	+0.107	−0.517	+0.279	+0.285

Table 5: Mean zero-shot bias (predicted − true index) by century, all six audited models, identical 702-row sample. "GPT-5" = GPT-5-mini; "Qwen" = Qwen3.7-Plus; "Kimi" = Kimi-K2.6-Fast.

Model	Slope/century	r	p
Gemini 2.5 Flash	+0.344	0.212	<0.0001
GPT-5-mini	+0.272	0.199	<0.0001
DeepSeek v4 Flash	+0.250	0.177	<0.0001
Claude Sonnet 4.6	+0.129	0.105	0.0054
Qwen3.7-Plus	+0.201	0.158	<0.0001
Kimi-K2.6-Fast	+0.257	0.199	<0.0001

Table 6: Linear regression of zero-shot bias against calendar year, $n = 702$ for all six models.

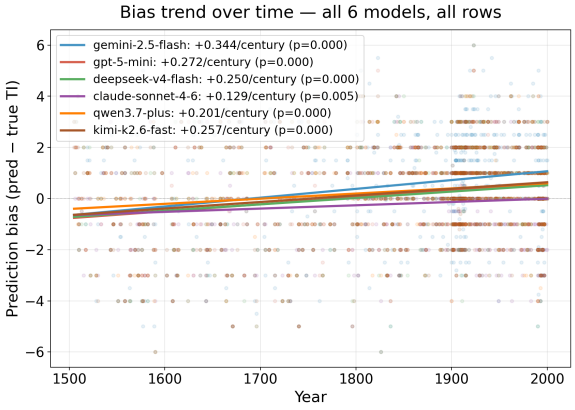


Figure 3: Bias-vs-year trend, all six audited models, $n = 702$ each (see Figure 1 for the corresponding per-century breakdown).

ing the temporal drift is largely decoupled from a model's baseline calibration quality. As a case study on Gemini, splitting the regression by the true index's sign shows the effect is strongest on cold months (slope $+0.389$/century, $p = 0.0007$): Gemini under-predicts cold severity progressively more for more recent centuries. It does not simply shift its mean prediction upward across the board.

We repeat this audit on the era-cue-stripped variant of the same 702 quotes described in §4 (explicit years and calendar-era markers removed). Table 7 shows the century-bias slope is essentially unchanged for every model: Gemini's slope moves from $+0.344$ to $+0.354$/century, GPT-5-mini's from $+0.272$ to $+0.269$, DeepSeek's from $+0.250$ to $+0.303$, and Claude's from $+0.129$ to $+0.124$. All remain significant ($p < 0.01$, three at $p < 0.0001$), with 96–122% of the original slope surviving redaction. Since the model can no longer read an explicit date or era marker off the text, this result points to an **anachronistic present-day prior** operating largely independently of the textual date evidence that was removed. The model does not appear to be simply reading the quote's era correctly from context. We discuss two residual, unredacted era cues that could still contribute part of this surviving slope in §6.

Two secondary checks probe whether the lexical-beats-transformer pattern from RQ1 generalizes beyond temperature. On the drought/humidity index, the same lexical baselines reach test correlations of 0.31–0.35, comparable to temperature's 0.28–0.32 (best MAE 1.19, char XGBoost, naturally higher given HHI's wider $[-4, +4]$ scale). On the pre-1500 medieval corpus, resolved seasonally instead of monthly, word-level TF-IDF reaches a substantially higher test correlation ($r = 0.4916$) despite far less training data (477 rows) and sparser, mul-

Model	Slope (orig.)	Slope (redacted)	p (red.)
Gemini 2.5 Flash	+0.344	+0.354	<0.0001
GPT-5-mini	+0.272	+0.269	<0.0001
DeepSeek v4 Flash	+0.250	+0.303	<0.0001
Claude Sonnet 4.6	+0.129	+0.124	0.0078

Table 7: Bias-vs-year regression slope (points/century), original vs. era-redacted text, $n = 702$ in both conditions.

tilingual source text. This is a surprising counterpoint to the intuition that older, sparser text carries a weaker signal, though the smaller test set (219 vs. 502 rows) makes this only suggestive.

As a structurally different test of where historical signal lives, we extend the same zero-shot/few-shot comparison to **event extraction on the medieval corpus**. Three in-context examples roughly double Gemini's strict-match F1 from 0.1855 to 0.3834 and cut year error from 6.7 to 2.8 years, while codebook adherence stays above 99.7% in both conditions even without schema enforcement. Contrary to the intuition that standardized Latin should be easier than fragmented medieval German dialects, node-label F1 is consistently *lower* on Latin quotes, both zero-shot (0.3204 vs. 0.4142) and few-shot (0.4486 vs. 0.4989).

6 Discussion

These results are connected. Simple lexical retrieval is hard to beat for extracting a climate signal from centuries-old text (RQ1). Prompting a general-purpose LLM with a handful of relevant examples can match or exceed even the best fine-tuned historical-domain transformer on correlation (RQ2). But zero-shot LLM errors follow a clear directional pattern: they run systematically warmer for more recent text, and this specific pattern replicates across six independently developed model families (RQ3). Together, RQ2 and RQ3 explain why a single correlation number is not enough on its own: a model can tie a lexical baseline on r (RQ2) while its error is concentrated on calendar year (RQ3), the exact dimension a historian or paleoclimatologist would need to trust for cross-century comparison.

We read the **warm-bias** trend mechanistically as an instance of the parametric-versus-contextual knowledge conflict described in §2 (Longpre et al., 2021; Xu et al., 2024). As there, the quote's date and content are the context, and the model's trained-in prior (§2) overrides that context once it can roughly infer the era from vocabulary or explicit dates. This connects to the presentism/anachronism literature (Underwood et al., 2025; Levchenko, 2025). Furthermore, it highlights a broader challenge in developing models capable of adapting to diverse perspectives (Creangă and Dinu, 2024), in this case, requiring the model to adopt a temporally distinct worldview rather than defaulting to a contemporary one. Our contribution is a numeric measurement of this failure mode's cost against real, expert-derived ground truth. Correctly inferring the era and applying a reasonable prior would produce the same bias signature as drifting toward a present-day prior regardless of context, but the two carry very different implications. Our era-cue-stripping ablation (§5) favors the latter: the slope survives, essentially unchanged, once explicit dates and calendar-era markers can no longer be read off the text.

7 Conclusion

We set out to test whether NLP and LLM systems can recover a physically meaningful climate signal from centuries of historical German text, and, if so, whether that signal is trustworthy enough for real cross-era scientific comparison. The answer to the first is yes: lexical methods extract this signal more reliably than every fine-tuned historical-domain transformer we test, and a general-purpose LLM can match the best lexical correlation with a handful of in-context examples. The answer to the second is more cautionary: all six LLMs we test show a systematic, era-correlated **warm bias that correlation alone does not reveal**. This bias survives an era-cue-stripping ablation run, favoring an anachronistic prior over the model correctly reading the quote's stated era.

Future work should test whether this bias generalizes to other regions and languages. Also, we can test whether human annotators asked to perform the same task show a comparable presentist bias, which would suggest the model is reproducing a documented human cognitive tendency rather than exhibiting an AI-specific failure mode. Relatedly, we can check whether part of the observed drift reflects semantic change in weather vocabulary over five centuries, such as what counted as a severe frost or a mild winter shifting between 1550 and 1950, rather than the model misjudging the quote's era itself.

Limitations

This work has several limitations. Our anachronistic-bias audit is statistically significant but explains a modest share of total prediction error ($r^2 \approx 0.01$–0.05 for the century-vs-bias trend across the six models, against an overall zero-shot MAE around 1.3–1.9 on a seven-point scale), and should not be read as the dominant source of error in LLM zero-shot climate-index prediction. The pre-1900 and post-1900 portions of our audit sample were constructed differently: one freshly stratified and sign-balanced, the other an older, unbalanced batch reused for cost efficiency. This is a sampling asymmetry we control for via sign-matched subsets but do not fully eliminate. Our era-cue-stripping ablation (§5) weighs against the observed bias reflecting a defensible era-appropriate prior. But two residual, unredacted era cues (two-digit year abbreviations and the structured 19th/20th-century statistical-report format, §6) could still contribute part of the surviving slope. We did not attempt automatic proper-noun redaction, since a quick manual check showed current historical-German NER handles this archaic orthography too poorly to trust. We test six LLM families, which we consider enough, but it doesn't confirm that the effect holds for all current LLMs. Older text in our corpus is also objectively harder to parse than modern text for reasons unrelated to anachronistic reasoning (denser archaic spelling, occasional Latin and French). This confound is hard to cleanly separate from our proposed mechanism. Finally, all data and evaluation are limited to German- (and to a lesser extent French- and Latin-) language historical sources from Central Europe. Whether the same pattern holds for other languages, regions, or climate-index conventions is untested.

Ethics Statement

This work uses historical documentary text and derived climate indices from the *tambora.org* research collaboratory (Riemann et al., 2015), a publicly available academic dataset. The source material and indices describe historical weather and climate conditions and contain no personal or sensitive information about identifiable individuals. All LLM queries were made through providers' standard commercial or research APIs under their respective terms of service, and involved no human subjects. We release no new personal data. We do not anticipate direct ethical risks from this work.

Acknowledgments

This work was partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416 and a grant of the Romanian Ministry of Research, Innovation and Digitalization, CNCS - UEFISCDI (Center of Excellence for Climate and Societal Change), project number PN-IV-P6-6.1-CoEx-2024-004, within PNCDI IV.

References

Anthropic. 2026. Claude Sonnet 4.6 system card. Technical report, Anthropic.

Christin Beck and Marisa Köllner. 2023. GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 33–45, Singapore. Association for Computational Linguistics.

Claudiu Creangă and Liviu P. Dinu. 2024. Designing NLP systems that adapt to diverse worldviews. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP*.

Claudiu Creanga, Liviu P. Dinu, and Anca Dinu. 2026. The Rashomon Wikipedia: A data-perspectivist analysis of divergent historical narratives. In *NLPerspectives@LREC 2026*, pages 11–20.

Miriam Cuscito, Alfio Ferrara, and Martin Ruskov. 2024. How BERT speaks Shakespearean English? Evaluating historical bias in masked language models. In *Proceedings of the 3rd Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2024)*, volume 3865 of *CEUR Workshop Proceedings*, Bolzano, Italy. CEUR-WS.org.

Miriam Cuscito, Alfio Ferrara, and Martin Ruskov. 2025. Shakespeare did not know our vocabulary: Measuring the historical adequacy of LLMs. In *2025 IEEE International Conference on Cyber Humanities (IEEE CH)*, Florence, Italy. IEEE.

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Rüdiger Glaser and Michael Kahle. 2020. Reconstructions of droughts in Germany since 1500 – combining hermeneutic information and instrumental records in historical and modern perspectives. *Climate of the Past*, 16(3):1207–1226.

Rüdiger Glaser and Dirk Riemann. 2009. A thousand-year record of temperature variations for Germany and Central Europe based on documentary data. *Journal of Quaternary Science*, 24(4):437–449.

Daniel Hutchinson. 2024. Mapping the latent past: Assessing large language models as digital tools through source criticism. *Journal of Digital History*, 3(1).

Kimi Team. 2026. Kimi K2.6: Advancing open-source coding.

Maria Levchenko. 2025. Building historical corpora with multimodal LLMs: Epistemic gaps and misreadings in 18th-century Russian books. In *Computational Humanities Research 2025*, volume 3 of *Anthology of Computers and the Humanities*, pages 725–737.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, Vancouver, Canada. PMLR.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ehsan Nabavi, Holger R. Maier, Saman Razavi, Adrian Hindes, Mark Howden, Will Grant, and Sujatha Raman. 2026. On using large language models to support social research for climate action. *npj Climate Action*, 5:52.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2026. GPT-5 system card. Technical report, OpenAI.

Christian Pfister. 1984. The potential of documentary data for the reconstruction of past climates: Early 16th to 19th century Switzerland as a case study. In *Climatic Changes on a Yearly to Millennial Basis*. Springer.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's high time: A survey of temporal question answering. *arXiv preprint arXiv:2505.20243*.

Qwen Team. 2026. Qwen3.7-Plus: Multimodal agent intelligence.

Dirk Riemann, Rüdiger Glaser, Michael Kahle, and Steffen Vogt. 2015. The CRE tambora.org – new data and tools for collaborative research in climate and environmental history. *Geoscience Data Journal*, 2(2):63–77.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. *arXiv preprint arXiv:2205.15575*.

Ted Underwood, Laura K. Nelson, and Matthew Wilkens. 2025. Can language models represent the past without anachronism? *arXiv preprint arXiv:2505.00030*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Khurram Yamin, Gaurav R. Ghosal, and Bryan Wilder. 2025. Can LLMs reconcile knowledge conflicts in counterfactual reasoning? *arXiv preprint arXiv:2506.15732*.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15015–15040, Bangkok, Thailand. Association for Computational Linguistics.

Improving Efficiency of Ternary Large Language Models through Pruning

Pietro Tropeano
University of Copenhagen
Copenhagen, Denmark
pitr@di.ku.dk

Maria Maistro
University of Copenhagen
Copenhagen, Denmark
mm@di.ku.dk

Tuukka Ruotsalo
University of Copenhagen
Copenhagen, Denmark
LUT University / Lahti, Finland
tr@di.ku.dk

Christina Lioma
University of Copenhagen
Copenhagen, Denmark
c.lioma@di.ku.dk

Nikolaos Aletras
University of Sheffield
Sheffield, United Kingdom
n.aletras@sheffield.ac.uk

Abstract

As the environmental impact of Large Language Models (LLMs) grows, improving model efficiency has become critical for sustainable NLP. Recent advances in model compression have shown that ternary quantisation (i.e., using only {-1, 0, 1} as weights) substantially reduces memory footprint, inference time, and energy usage of LLMs with minimal performance degradation. However, whether these already efficient models can be further pruned to reduce resource requirements without compromising effectiveness remains unexplored. We investigate this question across six models, four sparsity levels, and six structured and unstructured pruning techniques. We find that even minimal pruning yields substantial efficiency gains: pruning fewer than 5% of parameters reduces inference time and energy consumption by up to 21%, and memory usage by up to 13%, with minimal accuracy degradation. We also find that, for pruned models with less than a 5% accuracy drop, the most energy and inference-time efficient configurations are achieved by pruning entire attention layers, whereas the most memory efficient configurations are obtained by pruning entire MLP layers. These findings highlight pruning as a complementary technique to ternary quantisation for reducing environmental footprint of LLMs while maintaining model effectiveness. The code is available at: https://github.com/pietrotrope/TernaryLLMsPruning.

1 Introduction

Large Language Models (LLMs) offer impressive performance across a wide range of tasks (team, 2026; Liu et al., 2025; Yang et al., 2025a; Mistral-AI, 2025). However, they require significant computational resources when deployed. To mitigate this, model compression techniques such as quantisation and pruning are widely used (Dettmers et al., 2024; Frantar and Alistarh, 2023). Quantisation reduces the number of bits used to represent model parameters (Lang et al., 2024) and pruning removes parts of the model to reduce its size (Zhu et al., 2024). Both methods can be applied during or after training (Dettmers et al., 2024; Ma et al., 2024; Frantar and Alistarh, 2023; Kurtic et al., 2022), yet, when pruning LLMs, post-training pruning is the most common solution due to high training costs.

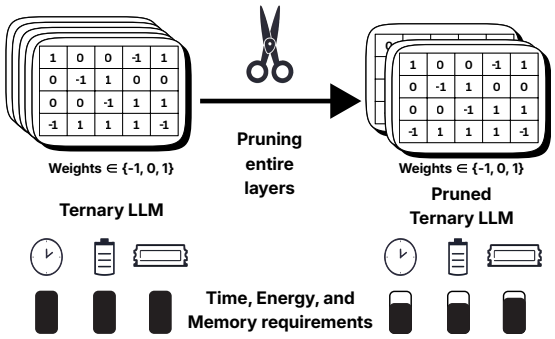


Figure 1: Efficiency gains of pruning ternary models. Pruning a small amount of entire LLMs' MLP or attention layers reduces time, energy, and memory usage without substantially compromising their accuracy.

Ternary quantisation such as the BitNet family of ternary models (Ma et al., 2024; Wang et al., 2023), uses only three values (i.e., $\{-1, 0, 1\}$) to represent weights. This extreme quantisation technique achieves accuracy comparable with similarly sized full precision models while requiring only

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 51–84
October 28, 2026 ©2026 Association for Computational Linguistics

1.58 bits per weight. Ternary models are designed for deployment in edge devices and low-resource settings (Ma et al., 2024), enabling broader accessibility and more environmentally sustainable use compared to full-precision models.

Although ternary quantisation reduces inference time and memory footprint, it does not reduce the number of model parameters. Pruning can complement quantisation by removing parameters directly, potentially further reducing memory usage, inference time, and energy consumption. Thus, pruning is relevant for sustainable NLP, as improving LLMs efficiency will reduce their environmental impact.

To this end, we contribute the first comprehensive analysis of post-training pruning for ternary LLMs. We study the extent to which ternary LLMs can be pruned without substantially compromising their effectiveness, and we characterise the resulting trade-offs between effectiveness and efficiency. Our study evaluates six pruning algorithms applied to six ternary models across four sparsity levels, measuring both downstream performance on seven datasets and the models' reduction in inference time, energy, and memory requirements.

We find that (i) layer pruning is an effective solution to prune ternary LLMs, allowing to remove a small number of parameters without substantially compromising accuracy. (ii) Even modest pruning, especially when removing entire layers, improves ternary LLMs' efficiency (Fig. 1). (iii) Layer pruning results in the best accuracy-efficiency trade-offs. Among pruned LLMs with limited accuracy degradation, pruning entire attention layers produces the best improvements in inference time and energy efficiency, whereas pruning entire MLP layers yields the greatest memory savings. Overall our findings highlight new opportunities to improve efficiency of low-resource LLMs.

2 Related Work

Model quantisation is a common compression approach that reduces the number of bits used to represent parameters (Lang et al., 2024), leading to practical memory and inference time efficiency improvements. While common quantisation methods reduce bits to 8, 4, or 2 (Xiao et al., 2023; Frantar et al., 2023; Chee et al., 2023), recent work, such as BitNet (Ma et al., 2024), has explored 1.58-bit models, also known as ternary models. These models allow only three possible states ($\{-1, 0, 1\}$), with accuracy that matches similarly sized full-precision models. Ternary models are natively sparse because one of their possible weight states is 0; the value conventionally used to prune weights.

Pruning is another compression technique that removes parameters or entire network structures according to a heuristic. Pruning algorithms can be unstructured or structured. Unstructured pruning removes individual parameters. Examples of such algorithms are Wanda (Sun et al., 2024), Ria (Zhang et al., 2024), and SparseGPT (Frantar and Alistarh, 2023). Structured pruning removes entire structures such as neurons or layers (Michel et al., 2019; Ma et al., 2023; Ashkboos et al., 2024). Unstructured pruning generally yields better accuracy-sparsity trade-offs (Hu et al., 2025a). Yet it does not offer inference speedups unless paired with custom hardware and sparsity patterns (Yang et al., 2025b). Structured pruning, on the other hand, provides tangible efficiency improvements by removing entire structures (Hu et al., 2025a). A hybrid approach, N:M semi-structured pruning, enables hardware acceleration by enforcing at most N non-zero parameters for every M contiguous parameters (Zhou et al., 2021). However, as current hardware accelerates only 2:4 sparse matrices, this limits pruning to remove exactly 50% of the weights and restricts more fine-grained sparsity exploration.

Pruning and quantisation are often combined, typically jointly, or by pruning first, then quantising the model (Han et al., 2016; van Baalen et al., 2020; Zhang et al., 2026). Zhang et al. (2026) applies semi-structured pruning during training of ternary LLMs. Yet, to the best of our knowledge, post-training pruning of ternary LLMs is unexplored, and addressing this gap is the focus of our work.

3 Study Design

To systematically analyse the interplay between extreme quantisation and sparsity, we design a comprehensive evaluation framework that subjects ternary LLMs to varying degrees of pruning. Our aim is to study how pruning ternary LLMs affects both *effectiveness* and *efficiency*.

3.1 Ternary Models

We use six ternary LLMs, covering a wide range of sizes (from 1B, up to 8B), and all four LLM families using the BitNet architecture (Wang et al., 2023). Specifically, we use BitNet 2B (Ma et al., 2024), Llama 3 8B 1.58b (Mekkouri et al., 2024), the 1B and 3B Falcon Edge models (Falcon-LLM

Team, 2025), and the 3B and 7B versions of Falcon 3 1.58b (Falcon-LLM Team, 2024).

3.2 Pruning Methods

We compress ternary LLMs using six widely-used algorithms covering three families of pruning approaches: (1) Unstructured pruning, which removes individual parameters; (2) Neuron pruning, which eliminates entire neurons (corresponding to rows in the weight matrices); and (3) Layer pruning, which removes entire layers from the network. We chose these algorithms to prune ternary models at different granularities, ranging from individual parameters to neurons to entire layers.

Unstructured Pruning. We prune individual parameters using two of the best-performing unstructured pruning algorithms for LLMs: Wanda (Sun et al., 2024) and Ria (Zhang et al., 2024). We exclude unstructured pruning methods that rely on weight reconstruction, i.e., algorithms that modify the remaining weights after pruning to recover accuracy, such as SparseGPT (Frantar and Alistarh, 2023), because these updated weights would violate the ternary constraints. Wanda determines which parameters to prune based on the absolute magnitude of weights (0 or 1 in ternary models) and activations. Ria extends Wanda by incorporating both row-wise and column-wise normalised weight magnitudes, averaging these two scores and using activations to guide pruning. We use the standard configurations for both algorithms. As Sun et al. (2024) and Zhang et al. (2024), we compute activations with 128 samples of 2048 tokens from the C4 dataset (Raffel et al., 2020).

Neuron Pruning. To prune neurons, we use the algorithm from Hu et al. (2025a) on neurons, which is a straightforward approach to remove neurons without retraining or violating ternary constraints. The algorithm computes each neuron's score by summing its parameter scores (using Wanda's scoring function) and removes neurons with the lowest scores. When a neuron is removed from an MLP block, the corresponding weights in subsequent linear layers are also pruned, eliminating the features produced by that neuron. For attention layers, we apply an analogous strategy by removing features that are consistently inactive, i.e., those that remain zero across 128 sequences of 2048 tokens. We use two variants for this method: one to prune attention layers and one for MLP. Since we work with ternary LLMs, we do not update weights after pruning, as in Hu et al. (2025a). Weights updates would result in non-ternary weights.

Dataset	Task	# samples	# Classes
ARC Easy	Multiple Choice QA	2376	4
ARC Challenge	Multiple Choice QA	1172	4
OpenBookQA	Multiple Choice QA	500	4
BoolQ	Question Answering (Yes/No)	3270	2
RTE	Natural Language Inference	277	2
Hellaswag	Natural Language Inference	10000	4
Winogrande	Commonsense reasoning	1267	2
Wikitext-2	Language Modelling	4358	-

Table 1: Datasets' task, number of samples, and classes.

Layer Pruning. For layer pruning, we use the lightweight method proposed by He et al. (2026). We select this approach because, unlike other methods, it does not yield domain-specific models (Hu et al., 2025b), and does not require further training (Lu et al., 2024). This pruning algorithm iteratively removes layers from an LLM based on their relative importance. Layer importance is estimated as the cosine similarity between its input and output representations (Men et al., 2025); where lower similarity implies higher importance. This relies on a small calibration set to obtain layer inputs and outputs. Since He et al. (2026) demonstrate that their method is robust to the quantity of calibration data, we use eight sequences of 2048 tokens sampled from the C4 dataset. Similar to Neuron Pruning, we use one variant that removes attention layers and another that removes MLP layers.

3.3 Model Sparsity Setup

To fairly compare different pruners, all pruned models must achieve roughly the same final sparsity. Otherwise, applying the same pruning threshold, a hyperparameter that controls how many parameters are removed, can remove very different numbers of parameters across methods, making efficiency comparisons unfair. This is not straightforward, as each method uses a pruning threshold, but its interpretation differs across strategies. In unstructured pruning, the threshold identifies the percentage of parameters to remove per layer, regardless of whether some of these parameters are already zero (i.e., already pruned). In neuron pruning, it indicates the fraction of neurons to remove, ignoring the sparsity of the remaining neurons. In layer pruning, it defines the number of layers to remove, independent of their internal sparsity.

As a result, the pruning threshold does not directly determine final sparsity. For instance, in unstructured pruning (e.g., Wanda), neurons that are already sparser than the pruning threshold are left unchanged. When combined with other neu-

rons that are pruned, the average sparsity of the layer will exceed the pruning threshold. Practically, if one neuron is already 45% sparse and the pruning threshold is 30%, the neuron is not pruned, and its higher sparsity adds to the overall layer, pushing the final average sparsity above 30%.

Prior work typically evaluates pruning methods within a single pruning family (e.g., structured or unstructured) on dense models (He et al., 2026; Sun et al., 2024; Frantar and Alistarh, 2023), making sparsity standardisation unnecessary. In contrast, we study already sparse models using different pruning families; therefore, we need to consider the target sparsity. To address this, we first select thresholds for attention-layer pruning. For other methods, we perform a binary search to find thresholds that achieve the closest final sparsity, stopping once the target is reached within 1% tolerance. This ensures that all pruned models achieve roughly the same sparsity. We present the final sparsities and pruning thresholds in the Appendix (see Tabs. 3-8).

3.4 Effectiveness Evaluation

We use seven question-answering and natural language inference datasets: ARC Easy and ARC Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), BoolQ (Clark et al., 2019), RTE (Wang et al., 2018), Hellaswag (Zellers et al., 2019), and Winogrande (Sakaguchi et al., 2021). We measure 0-shot accuracy on the test set of all datasets, using LM Evaluation Harness (Gao et al., 2024) standard prompt templates. We also evaluate language modelling effectiveness by computing perplexity on the Wikitext-2 dataset (Merity et al., 2017). Tab. 1 presents an overview of the datasets. Except for layer pruning, which removes layers in a fixed order and produces identical models, all results are averaged over three random seeds.

3.5 Efficiency Evaluation

We evaluate efficiency by measuring memory, GPU time, and energy required by the models. Following (Ashkboos et al., 2024; Sun et al., 2024; Frantar and Alistarh, 2023), we use matrix multiplication time, which requires the most computation in LLMs, as a proxy for inference time. We extend this approach by tracking memory and energy consumption, and by measuring all layers individually rather than using a single representative matrix per layer type.

We compute efficiency as follows. For layer pruning, we exclude the computation time of removed layers and aggregate the inference time of the remaining layers. For neuron pruning, we instantiate smaller layers with the pruned neurons removed, measure the resulting matrix multiplication costs, and sum the times across all layers. All evaluations use real activations sampled from 128 sequences of 2048 tokens from the C4 dataset.

We compute memory savings based on the layer's total number of parameters (including zeros) and the number of bits used to represent them. GPU time and energy are tracked on full-precision matrices (i.e., 32 bits) using Carbontracker (Anthony et al., 2020), a framework that tracks resource usage of Deep Learning models.

We note that unstructured pruning approaches do not yield more efficient models directly, as they require dedicated hardware to leverage sparsity, which currently does not exist. Hence, we only track the efficiency of structured pruning algorithms. Furthermore, unlike Sun et al. (2024) and Frantar and Alistarh (2023), we exclude NVIDIA Cutlass acceleration for N:M sparsity (where a matrix admits at most N non-zero parameters for every contiguous M parameters). This is necessary as our pruned models are not N:M sparse, and enforcing this structure would constrain unstructured pruning to a 50% sparsity.

Accuracy-Efficiency trade-off. We study the accuracy-efficiency trade-off of our pruned LLMs using the measure Performance per Resource Unit (PePR) (Selvan et al., 2025). PePR is defined as:

$$PePR(R, P) = \frac{P}{1 + R} \tag{1}$$

where $R = (R - R_{min})/(R_{max} - R_{min})$ is the min-max normalised resource cost in the experimental setting. In our experiments, we track memory and energy consumption and apply normalisation separately for each model, considering its various pruned versions. P is a normalised measure of model performance, in our case, accuracy. The PePR score ranges between 0 and 1. Lower scores indicate a worse trade-off between effectiveness and efficiency, while a score of 0 corresponds to a model with 0% accuracy. A score of 1 indicates a model that achieves 100% accuracy while being the most efficient in terms of the chosen resource.

4 Results

4.1 Effectiveness

Figs. 2 (a-g) show the accuracy of pruned and unpruned BitNet across all datasets (see Figs. 6-10 in Appendix for other models). Across datasets and

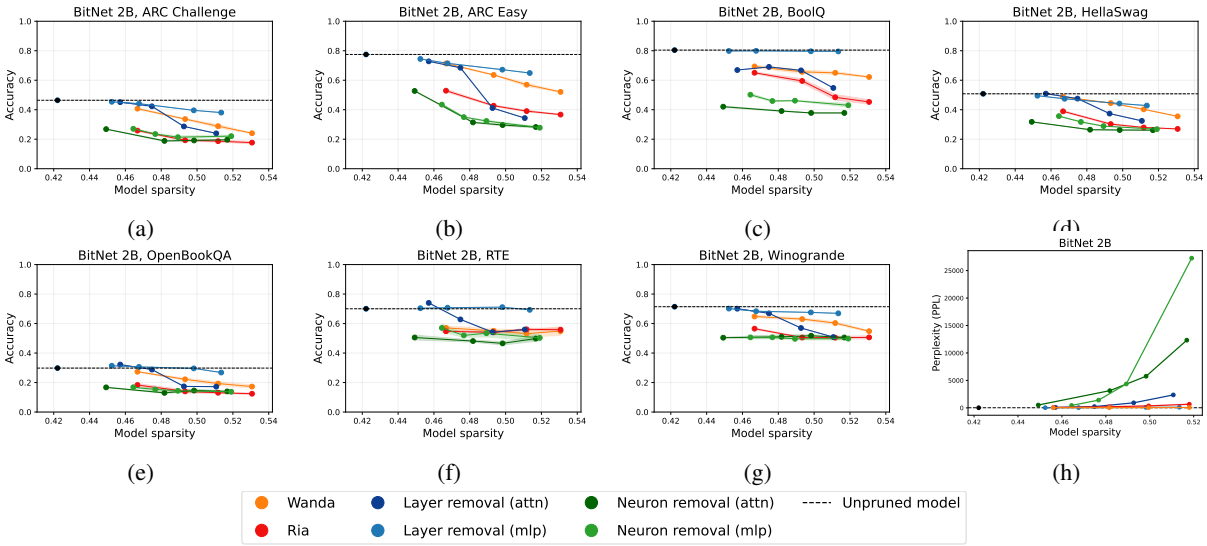


Figure 2: Panels (a)-(g): **Accuracy** of pruned and unpruned **BitNet** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

models, pruning MLP layers consistently yields the highest accuracy given a model sparsity. In contrast, neuron removal on attention layers results in the lowest accuracy across sparsity levels.

Overall, MLP layer pruning can increase model sparsity up to 5% while keeping accuracy drops below 5%. The only exception is Falcon-E 3B, for which our experimental setup limits the sparsity increases to 4%. Nevertheless, the same trend holds when increasing sparsity up to 4%. These results demonstrate that *ternary models, despite being inherently sparse, can be further sparsified through pruning with high accuracy retention*.

Fig. 2 (h) shows models' perplexity on Wikitext-2 (see Appendix, Figs. 6-10 for other models). Unstructured pruning (Wanda and Ria) achieve the lowest perplexity when pruning the most parameters, whereas neuron removal yields the highest perplexity. For both neuron and layer removal, pruning attention layers leads to higher perplexity than pruning MLP. This trend holds across models.

4.2 Efficiency

We next quantify the impact of pruning on ternary models efficiency by analysing their inference time, energy consumption, and memory requirements on 128 samples from the C4 dataset. We remove outliers that deviate by more than three standard deviations from the mean matrix multiplication time. These outliers account for less than 2.2% of the data and are sporadic measurements with abnormally large runtimes, which would skew the estimates.

Fig. 3 shows the average inference time (first column), energy consumption (second column), and memory requirements (third column) of pruned BitNet and LLaMA models at a fixed target sparsity (as the trends are consistent across models, the remaining models are provided in Appendix, Figs. 11, 12, and 13). Across all models, pruning attention layers (dark blue lines) consistently yields the lowest inference time for a given target sparsity.[1] In contrast, pruning entire neurons (green lines) in the attention and MLP layers results in slower inference in most settings. We hypothesise that this behaviour arises because GPUs are highly optimised for matrix operations with specific dimensions; removing neurons may increase computation time rather than reducing it if the shape of the new matrix is not aligned with GPU-optimised cores (NVIDIA, 2023). Similarly, when it comes to energy consumption (second column), except for Falcon-3 7B, removing attention layers produces the most energy-efficient models across sparsity levels. Last, for a fixed target sparsity, all pruning algorithms achieve similar reductions in memory footprint, except for BitNet, where MLP neuron removal leads to greater memory savings than the other pruners. Overall, this means that in our settings, *pruning attention layers is the most effective*

[1]This advantage would likely be greater in deployment, where pruning attention layers reduces KV-cache overhead; here we measure only matrix multiplication time.

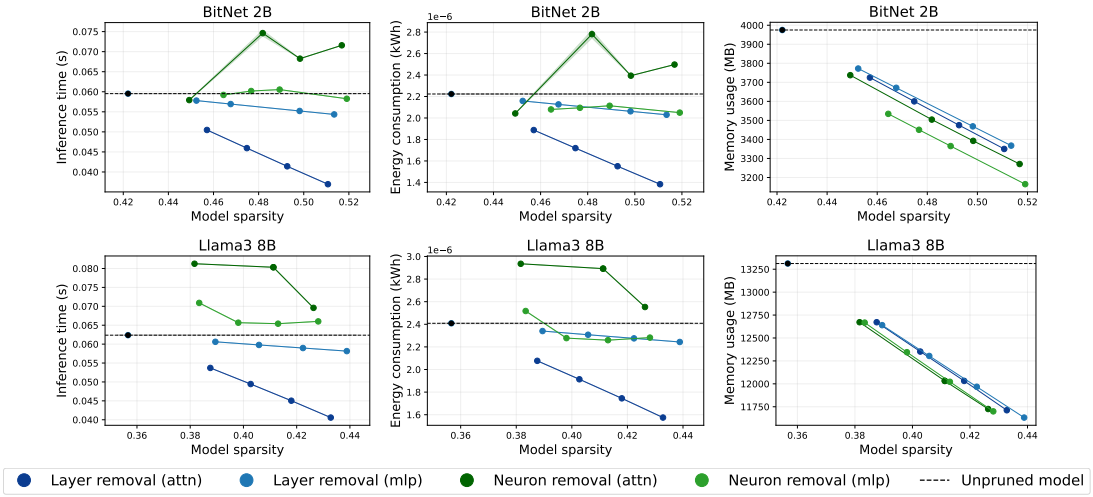


Figure 3: Efficiency of pruned models as a function of sparsity. The first, second, and third columns show inference time, energy, and memory, respectively, with sparsity on the x-axis. Coloured lines represent different pruning methods, while the black dotted line indicates the unpruned model. The black dot marks the sparsity of the unpruned model. For time and energy, all shaded regions indicate variability across 128 samples (± standard deviation).

strategy for reducing time and energy requirements.

4.3 Effectiveness-efficiency Trade-off

Next, we evaluate accuracy-efficiency trade-off of pruned models using the PePR score (see Sec. 3.4).

Performance Per Resource Unit. Fig. 4 shows the PePR time (first column), PePR energy (second column), and PePR memory (third column) scores of different pruned versions of BitNet as a function of their final sparsity on two datasets. Similar trends are observed across all datasets and model architectures (see Appendix, Figs. 14-31).

When pruning a small fraction of parameters, removing entire attention layers (dark blue lines) yields the best performance per unit of inference time and energy. However, as sparsity increases, pruning entire MLP layers (light blue lines) often produces models that require less time and energy to achieve the same performance. Across all sparsity levels, pruning MLP layers is the strategy that most frequently maximises performance per unit of memory. In contrast, neuron-level pruning (green lines) significantly harms model accuracy, consistently resulting in worse PePR scores than the original unpruned model (black dot).

Across all models and datasets, we consistently find pruned configurations that outperform the original model in time or memory efficiency while achieving higher PePR scores, indicating better accuracy per unit of time, memory, or energy.

Overall, pruning either MLP or attention layers yields the best accuracy-efficiency trade-offs.

Accuracy against Efficiency. Performance per resource unit alone does not provide a complete picture of a model's suitability for deployment, as it collapses accuracy and efficiency into a single scalar value. This aggregation does not show which model best satisfies specific deployment constraints, such as strict limits on memory usage or energy consumption. Given a concrete deployment bottleneck, a more informative approach is to directly examine model accuracy as a function of the constrained resource. Accordingly, we next study the accuracy-time, accuracy-energy, and accuracy-memory trade-offs of the pruned models.

Fig. 5 shows the accuracy of pruned BitNet models on two classification tasks against their mean inference time (first column), energy (second column), and memory requirements (third column). Similar trends hold across other datasets and model architectures (Appendix, Figs. 32-49).

Across all three efficiency aspects, in the best case, removing entire attention layers yields models that remain close in accuracy to the unpruned model (black point), with accuracy drops not exceeding 5% (except for BoolQ, where the drop reaches 11%). This pruning strategy increases sparsity by less than 4% while delivering substantial efficiency gains, including up to a 21% inference time and energy consumption reduction. Pruning MLP layers also preserves accuracy but results in more limited reductions in inference time and energy usage. These trends hold consistently across all models, except Falcon-3 3B, where removing at-

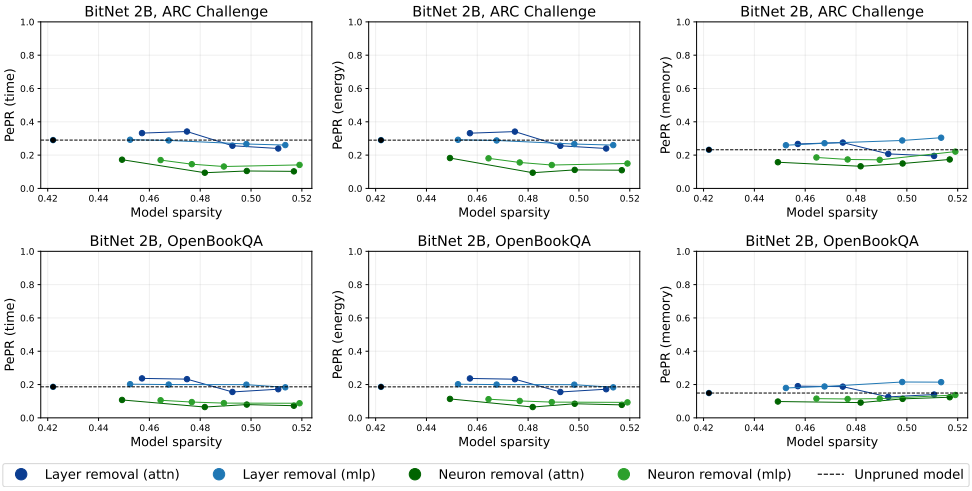


Figure 4: PePR score of pruned models as a function of sparsity. The first, second, and third columns show PePR scores for time, energy, and memory, respectively, with sparsity on the x-axis. Results on the first row are for ARC challenge, while results on the second row are for OpenbookQA. Coloured lines represent different pruners, while the black dotted line indicates the unpruned model. The black dot marks the sparsity of the original unpruned model.

Model	Pruning Algorithm	Sparsity	Accuracy	Accuracy Drop	Energy (1e5 Kwh)	Time (s)	Memory (MB)
BitNet 2B	Unpruned	0.422	0.609	-	0.222	0.060	3975
BitNet 2B	Layer removal (attn)	0.457	0.589	0.020	0.189	0.050	3725
BitNet 2B	Layer removal (mlp)	0.452	0.602	0.007	0.216	0.058	3773
BitNet 2B	Layer removal (mlp)	0.468	0.590	0.020	0.213	0.057	3671
BitNet 2B	Layer removal (mlp)	0.498	0.570	0.039	0.206	0.055	3469
Llama3 8B	Unpruned	0.357	0.543	-	0.241	0.062	13312
Llama3 8B	Layer removal (attn)	0.388	0.532	0.011	0.208	0.054	12672
Llama3 8B	Layer removal (attn)	0.403	0.510	0.033	0.191	0.049	12352
Llama3 8B	Layer removal (mlp)	0.389	0.532	0.010	0.234	0.061	12640
Llama3 8B	Layer removal (mlp)	0.406	0.515	0.028	0.231	0.060	12304
Llama3 8B	Layer removal (mlp)	0.422	0.506	0.036	0.228	0.059	11968

Table 2: Configurations of pruned BitNet 2B and Llama3 8B where the drop in accuracy does not exceed 5%.

tention layers always leads to substantial accuracy drops on ARC Easy and BoolQ.

In terms of memory efficiency, for accuracy drops of up to 5%, pruning MLP layers leads to the best memory-accuracy trade-off. Across the evaluated models, it reduces memory requirements by 5.6-13%. The largest reduction is achieved for Llama 8B, where memory usage decreases by more than 1.3 GB (13%) while maintaining an accuracy drop below 5% (see Appendix, Figs. 43, 49).

These results show that modest levels of pruning can yield substantial deployment benefits: in our settings, *pruning entire MLPs is the most effective approach to reduce memory footprint (13%), without substantially lowering accuracy, whereas pruning entire attention layers provides the largest reductions in inference time and energy consumption (21%) while maintaining competitive accuracy*.

5 Recommendations

Our experiments provide a comprehensive view of how pruning affects ternary models' accuracy and efficiency. Notably, we reveal consistent patterns across architectures, datasets, and resource constraints, offering practical deployment guidelines.

Pruning a small number of ternary models' parameters preserves accuracy while substantially improving efficiency (See Tab. 2 for examples on BitNet 2B and Llama3 8B). Pruning attention layers yields the largest reductions in inference time and energy consumption (21%), while pruning MLP layers achieves the greatest memory savings (13%), both without exceeding a 5% accuracy loss. For example, serving one million queries[2] with the most energy-efficient pruned BitNet model (removing eight attention layers) reduces estimated CO_2 emissions from 851 to 725 g in the USA, 764 to 650 g

[2]Individual queries served without batching, assuming an average sequence length of 2,048 tokens.

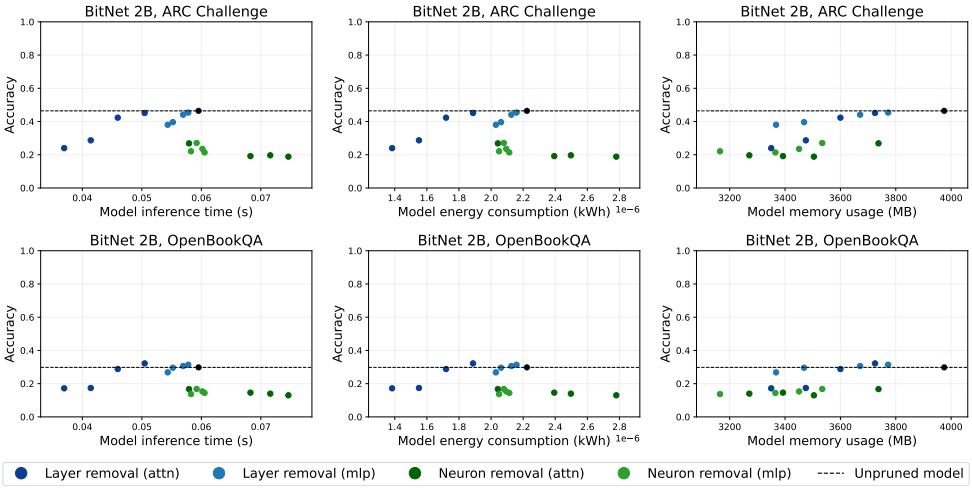


Figure 5: Efficiency of pruned models against model accuracy. The first, second, and third columns show inference time, energy, and memory, respectively, with accuracy on the y-axis. Results on the first row are for ARC challenge. Results on the second row are for OpenbookQA. Coloured dots represent different pruning methods, while the black line indicates unpruned model's accuracy. Black dot marks time, energy, and memory for the unpruned model.

in Germany, and 1,242 to 1,058 g in China. For the USA, this reduction is equivalent to decreasing the emissions associated with driving a passenger car from approximately 8.0 km to 6.8 km.[3]

Since entire layers are pruned, we expect efficiency improvements to generalise to inference frameworks such as BitNet.cpp (Wang et al., 2024), which provides optimised kernels for inference with packed ternary models (i.e., models that store multiple ternary weights per byte). In contrast, the other pruning methods would not reduce the memory footprint of packed ternary models. For instance, neuron pruning degrades accuracy, offers poor efficiency trade-offs, and is incompatible with BitNet.cpp, as it alters matrix sizes and breaks optimised kernels, making it unsuitable for low-resource deployment. Unstructured pruning cannot currently yield efficiency gains on low-resource hardware either, as it requires strict N:M sparsity patterns and specialised hardware to improve efficiency. Overall, when balancing efficiency and accuracy, attention layer pruning is the best strategy to minimise time and energy constraints, whereas MLP layer pruning is best for memory constraints.

Lastly, when pruning ternary LLMs, we recommend (i) pruning entire attention or MLP layers first, depending on the deployment constraint. (ii) Increase sparsity moderately (typically below 5%) to maintain accuracy while gaining efficiency. (iii) Evaluate multiple pruning configurations using both PePR scores and effectiveness-efficiency models comparison to find the optimal trade-off for the target deployment environment.

[3] CO_2 emissions and driving distances are estimated using country-specific emission factors from CarbonTracker.

6 Conclusions

In this work, we studied whether already compact ternary LLMs can be pruned to improve memory footprint, inference time, and energy consumption, while preserving effectiveness. We evaluated six pruning algorithms across six ternary LLMs and seven datasets, analysing accuracy, GPU inference time, energy consumption, and memory footprint. We found that layer pruning can remove a moderate number of parameters (e.g., 5%) from ternary LLMs, improving their efficiency without substantially compromising accuracy. Ternary LLMs are more sensitive to pruning than reported for full-precision models. Moreover, unlike prior work on full-precision models (Hu et al., 2025a), layer pruning yielded the highest accuracy in our experiments. Importantly, pruning specific layer types can help target deployment bottlenecks while preserving accuracy. Across models, the best memory trade-offs within a 5% accuracy drop were achieved by pruning MLP layers, while the best inference time and energy trade-offs were obtained by pruning attention layers, providing practical guidance for deployment in resource-constrained environments.

Future work will explore pruning approaches that combine attention and MLP layer removal to better understand their effectiveness-efficiency trade-offs. Additionally, pruning methods specif-

ically tailored to ternary models should be developed, including strategies for weight reconstruction that respect ternary constraints, as common methods like SparseGPT cannot be directly applied. Further work will also explore the end-to-end decoding efficiency of pruned ternary models on specific inference frameworks and formats such as BitNet.cpp

Limitations

This study investigates the effectiveness-efficiency trade-offs of pruning ternary models using unstructured, neuron, and layer pruning. While we cover a variety of pruning strategies, additional heuristics within each family could be explored. We evaluated efficiency by tracking GPU time, energy, and memory for matrix multiplications, but other approaches could produce different results, such as measuring CPU resource usage or efficiency in alternative model formats like BitNet.cpp. As our experiments focus on ternary models, the findings may not directly generalise to full-precision models, though studying their effectiveness-efficiency trade-offs would be valuable. Finally, we only evaluated ternary LLMs based on the BitNet architecture (Wang et al., 2023), and further research is needed for other ternary models, such as the Spectra suite (Kaushal et al., 2024).

Ethics Statement

This study focuses on investigating the effectiveness-efficiency trade-offs of pruning ternary LLMs to support future deployment in resource-constrained environments. While pruning improves efficiency and accessibility, these models remain subject to limitations and potential risks, including biases, hallucinations, misinformation, or inappropriate outputs. Models should be deployed responsibly with appropriate safety measures in place. All experiments were conducted on publicly available datasets and models, without using private or sensitive data. We will share our code (upon acceptance) to support reproducibility and further research.

Acknowledgments

This work was funded by the Algorithms, Data & Democracy project (Villum & Velux foundations). We also thank the IR Lab of the University of Copenhagen, in particular, Theresia Veronika Rampisela and Shivam Adarsh for their feedback on a draft of this paper. Lastly, we thank the anonymous reviewers for their valuable comments and suggestions that helped improve earlier versions of the paper.

References

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. *Preprint*, arXiv:2007.03051.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. SliceGPT: Compress Large Language Models by Deleting Rows and Columns. In *International Conference on Learning Representations*, volume 2024, pages 11682–11701.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. 2023. QuIP: 2-Bit Quantization of Large Language Models With Guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pages 4396–4429. Curran Associates, Inc.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Tim Dettmers, Ruslan A Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2024. SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression. In *The Twelfth International Conference on Learning Representations*.

Falcon-LLM Team. 2024. The Falcon 3 Family of Open Models.

Falcon-LLM Team. 2025. Falcon-E, a series of powerful, universal and fine-tunable 1.58bit language models.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations* .

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The Language Model Evaluation Harness.

Song Han, Huizi Mao, and William J. Dally. 2016. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. 2026. Uncovering the Redundancy in Transformers via a Unified Study of Layer Dropping. *Transactions on Machine Learning Research*.

Hanyu Hu, Pengxiang Zhao, Ping Li, Yi Zheng, Zhefeng Wang, and Xiaoming Yuan. 2025a. FASP: Fast and Accurate Structured Pruning of Large Language Models. *Preprint*, arXiv:2501.09412.

Lanxiang Hu, Tajana Rosing, and Hao Zhang. 2025b. TrimLLM: Progressive Layer Dropping for Domain-Specific LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 667–681, Vienna, Austria. Association for Computational Linguistics.

Ayush Kaushal, Tejas Vaidhya, Arnab Kumar Mondal, Tejas Pandey, Aaryan Bhagat, and Irina Rish. 2024. Spectra: Surprising effectiveness of pretraining ternary language models at scale. *arXiv preprint arXiv:2407.12327*.

Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. 2022. The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4163–4181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiedong Lang, Zhehao Guo, and Shuyu Huang. 2024. A Comprehensive Study on Quantization Techniques for Large Language Models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yao Lu, Hao Cheng, Yujie Fang, Zeyu Wang, Jiaheng Wei, Dongwei Xu, Qi Xuan, Xiaoniu Yang, and Zhaowei Zhu. 2024. Reassessing Layer Pruning in LLMs: New Insights and Methods. *arXiv preprint arXiv:2411.15558*.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Lifeng Dong, Ruiping Wang, Jilong Xue, and Furu Wei. 2024. The era of 1-bit LLMs: All large language models are in 1.58 bits. *Preprint*, arXiv:2402.17764.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-Pruner: On the Structural Pruning of Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 21702–21720. Curran Associates, Inc.

Mohamed Mekkouri, Marc Sun, Leandro von Werra, and Thomas Wolf. 2024. 1.58-Bit LLM: A New Era of Extreme Quantization.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mistral-AI. 2025. Magistral. *Preprint*, arXiv:2506.10910.

NVIDIA. 2023. NVIDIA Docs: Matrix Multiplication Background User's Guide.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Raghavendra Selvan, Bob Pepin, Christian Igel, Gabrielle Samuel, and Erik B Dam. 2025. PePR: Performance Per Resource Unit as a Metric to Promote Small-scale Deep Learning. In *Proceedings of the 6th Northern Lights Deep Learning Conference (NLDL)*, volume 265 of *Proceedings of Machine Learning Research*, pages 220–229. PMLR.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2024. A Simple and Effective Pruning Approach for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

OpenAI team. 2026. OpenAI GPT-5 System Card. *Preprint*, arXiv:2601.03267.

Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. 2020. Bayesian Bits: Unifying Quantization and Pruning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5741–5752. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. 2023. Bitnet: Scaling 1-bit transformers for large language models. *Preprint*, arXiv:2310.11453.

Jinheng Wang, Hansong Zhou, Ting Song, Shaoguang Mao, Shuming Ma, Hongyu Wang, Yan Xia, and Furu Wei. 2024. 1-bit AI infra: Part 1.1, fast and lossless bitnet b1. 58 inference on cpus. *Preprint*, arXiv:2410.16144.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mingzhe Yang, Sihao Lin, Changlin Li, and Xiaojun Chang. 2025b. Let LLM tell what to prune and how much to prune. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 70833–70849. PMLR.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Di Zhang, Xun Wu, Shaohan Huang, Yudong Wang, Hanyong Shao, Yingbo Hao, Zewen Chi, Li Dong, Ting Song, Yan Xia, Zhifang Sui, and Furu Wei. 2026. Sparse-BitNet: 1.58-bit LLMs are Naturally Friendly to Semi-Structured Sparsity. *Preprint*, arXiv:2603.05168.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024. Plug-and-Play: An Efficient Post-training Pruning Method for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Aojun Zhou, Yukun Ma, Junnan Zhu, Jianbo Liu, Zhijie Zhang, Kun Yuan, Wenxiu Sun, and Hongsheng Li. 2021. Learning N:M Fine-grained Structured Sparse Neural Networks From Scratch. In *International Conference on Learning Representations*.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A Survey on Model Compression for Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

A Study Design

We present the pruning thresholds and final sparsities for all configurations of models and pruning algorithms in Tables 3-8.

All experiments are performed on NVIDIA H100, except for effectiveness, which is tracked using both A100 and H100 GPUs.

B Results

Figs. 6-10 show the accuracy of pruned and unpruned ternary models across all datasets.

Figs. 11, 12, and 13 show the average inference time, energy consumption, and memory requirements, respectively, at different sparsity levels for all pruned models.

Figs. 14 to 31 show the PePR time, energy, and memory, respectively, at different sparsity levels for all pruned models.

Figs. 32 to 49 show the inference time, energy consumption, and memory requirements, respectively, against accuracy for all pruned models.

Tabs. 9-14 show a summary of the accuracy, energy consumption, time, and memory measurements obtained for all pruned configurations of the models.

pruner	threshold	sparsity
Layer_removal (attn)	8	0.457
	12	0.475
	16	0.493
	20	0.511
Layer_removal (mlp)	2	0.452
	3	0.468
	5	0.498
	6	0.513
Neuron_removal (attn)	25	0.449
	50	0.482
	62	0.498
	75	0.517
Neuron_removal (mlp)	12	0.464
	15	0.477
	18	0.489
	25	0.519
Wanda	43	0.456
	46	0.482
	48	0.499
	50	0.518

Table 3: Pruning thresholds and final sparsities for bitnet-b1.58-2B-4T

pruner	threshold	sparsity
Layer_removal (attn)	8	0.437
	12	0.450
	16	0.464
	20	0.478
Layer_removal (mlp)	1	0.431
	2	0.452
	3	0.472
Neuron_removal (attn)	25	0.428
	50	0.448
	75	0.469
Neuron_removal (mlp)	6	0.438
	9	0.452
	12	0.466
	15	0.481
Wanda	43	0.443
	46	0.467
	48	0.485

Table 4: Pruning thresholds and final sparsities for Falcon-E-1B-Base

pruner	threshold	sparsity
Layer_removal (attn)	8	0.411
	12	0.418
	16	0.425
	20	0.433
Layer_removal (mlp)	1	0.413
	2	0.430
Neuron_removal (attn)	25	0.409
	50	0.423
Neuron_removal (mlp)	3	0.411
	6	0.426
Wanda	37	0.403
	40	0.415
	41	0.421
	43	0.437

Table 5: Pruning thresholds and final sparsities for Falcon-E-3B-Base

pruner	threshold	sparsity
Layer_removal (attn)	8	0.388
	12	0.403
	16	0.418
	20	0.433
Layer_removal (mlp)	2	0.389
	3	0.406
	4	0.422
	5	0.439
Neuron_removal (attn)	25	0.382
	50	0.411
	62	0.426
Neuron_removal (mlp)	6	0.383
	9	0.398
	12	0.413
	15	0.428
Wanda	37	0.385
	40	0.409
	43	0.437

Table 6: Pruning thresholds and final sparsities for Llama3-8B-1.58-100B-tokens

pruner	threshold	sparsity
Layer_removal (attn)	8	0.342
	12	0.371
	16	0.401
	20	0.430
Layer_removal (mlp)	2	0.333
	4	0.383
	5	0.409
	6	0.434
Neuron_removal (attn)	37	0.337
	62	0.378
	75	0.400
	88	0.423
Neuron_removal (mlp)	12	0.344
	18	0.376
	21	0.392
	28	0.430
Wanda	34	0.344
	37	0.372
	40	0.401
	43	0.430

Table 7: Pruning thresholds and final sparsities for Falcon3-3B-Base-1.58bit

pruner	threshold	sparsity
Layer_removal (attn)	8	0.325
	12	0.335
	16	0.345
	20	0.356
Layer_removal (mlp)	1	0.326
	2	0.349
Neuron_removal (attn)	25	0.320
	50	0.338
Neuron_removal (mlp)	3	0.320
	6	0.337
	9	0.355
Wanda	31	0.320
	34	0.345
	35	0.354

Table 8: Pruning thresholds and final sparsities for Falcon3-7B-Base-1.58bit

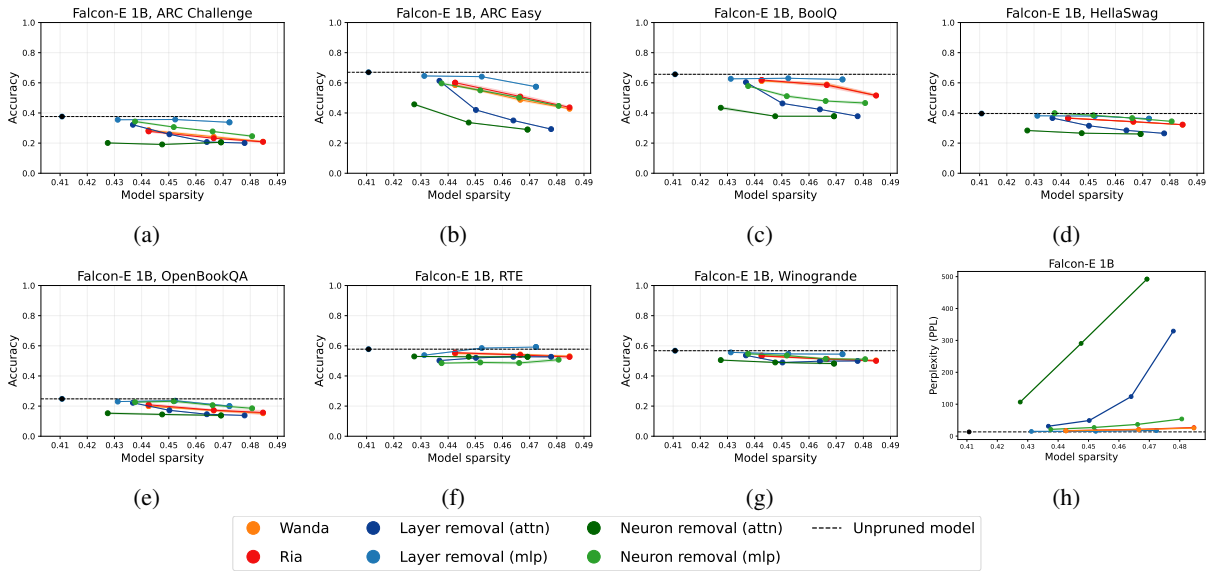


Figure 6: Panels (a)-(g): **Accuracy** of pruned and unpruned **Falcon-E 1B** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

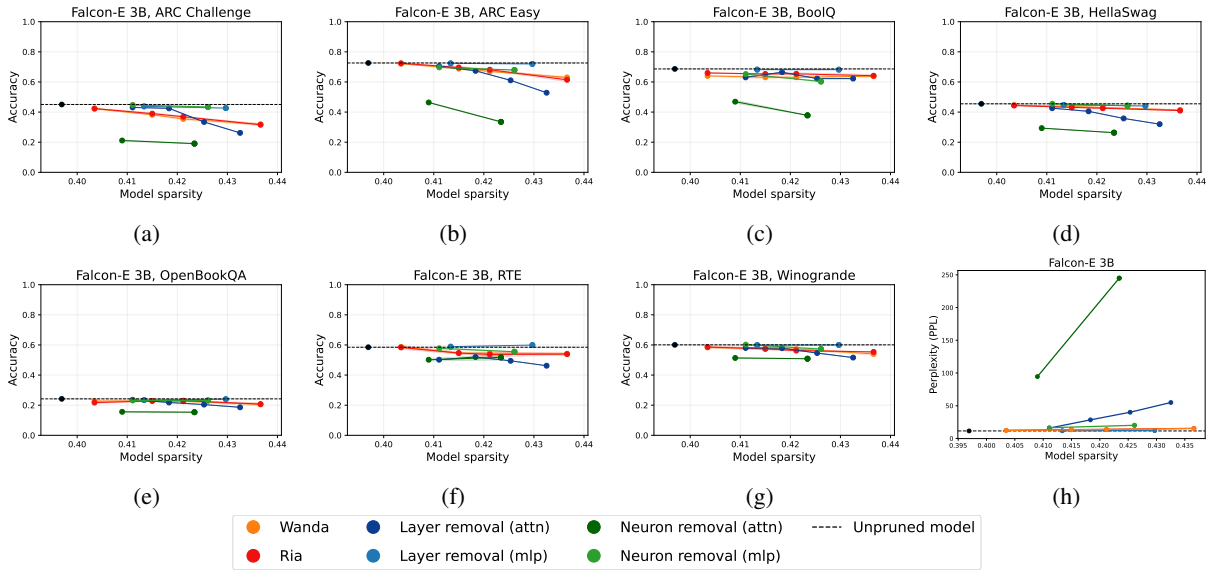


Figure 7: Panels (a)-(g): **Accuracy** of pruned and unpruned **Falcon-E 3B** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

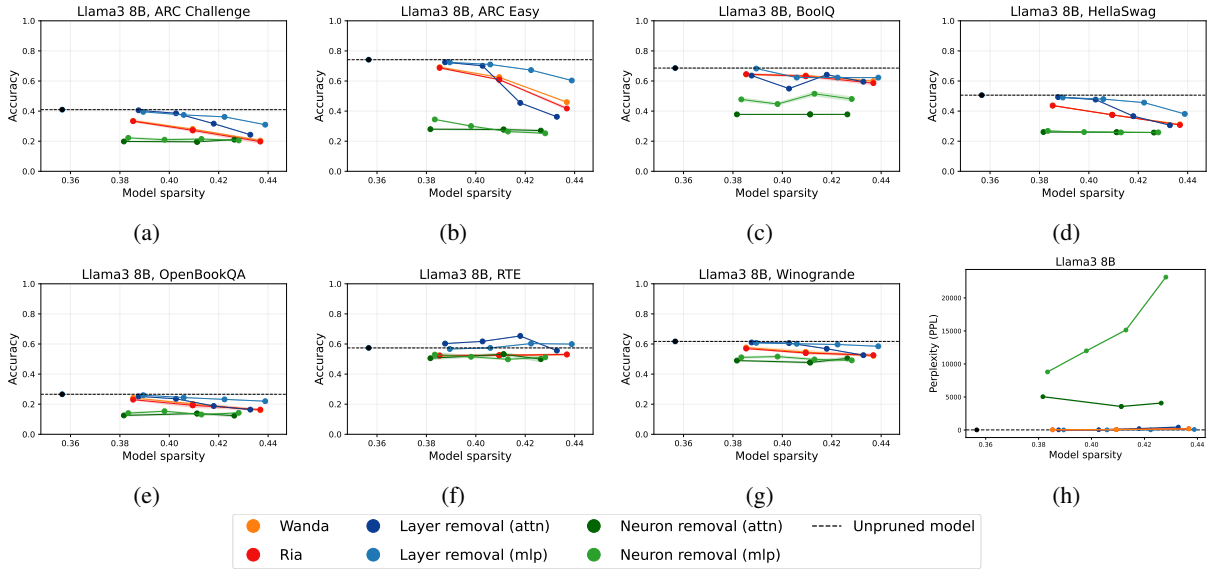


Figure 8: Panels (a)-(g): **Accuracy** of pruned and unpruned **Llama3** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

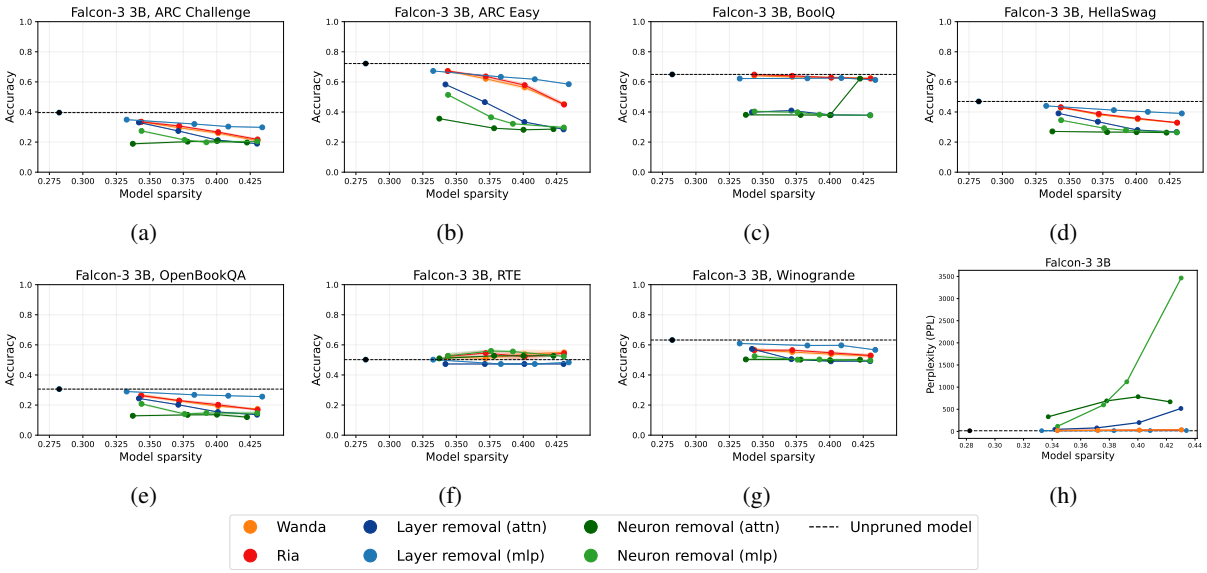


Figure 9: Panels (a)-(g): **Accuracy** of pruned and unpruned **Falcon3 3B** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

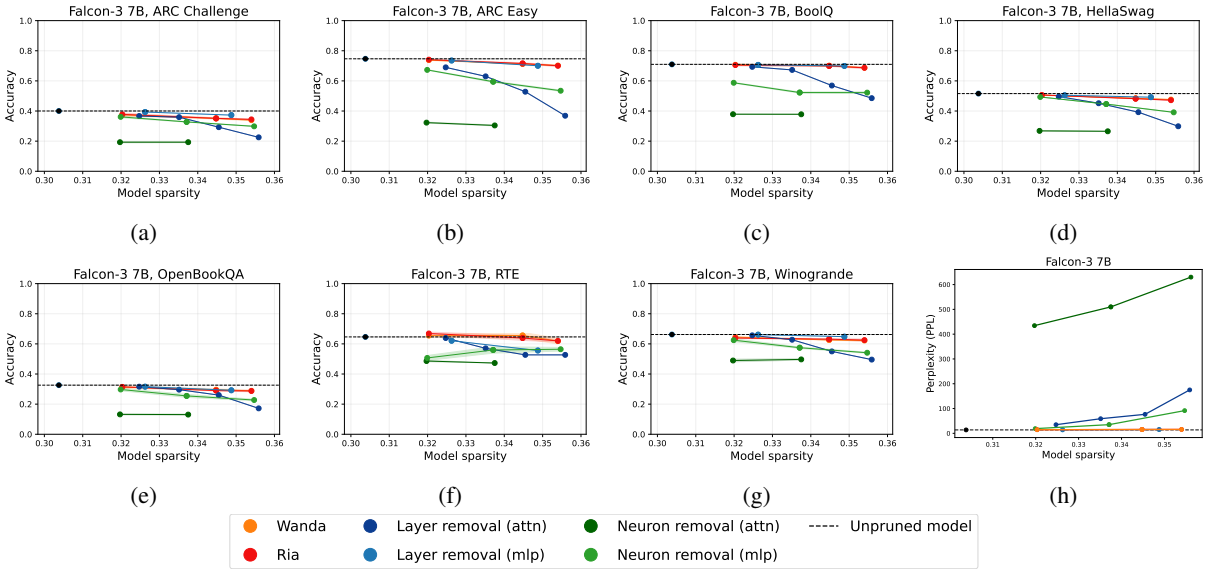


Figure 10: Panels (a)-(g): **Accuracy** of pruned and unpruned **Falcon3 7B** on seven datasets (y-axis), plotted against the **relative sparsity** of the pruned models compared to the original model (x-axis). Panel (h): **Perplexity** of pruned and unpruned BitNet on the Wikitext-2 dataset (y-axis) versus relative sparsity (x-axis). Unpruned models are shown with dashed lines. All shaded regions indicate variability across three runs (± standard deviation).

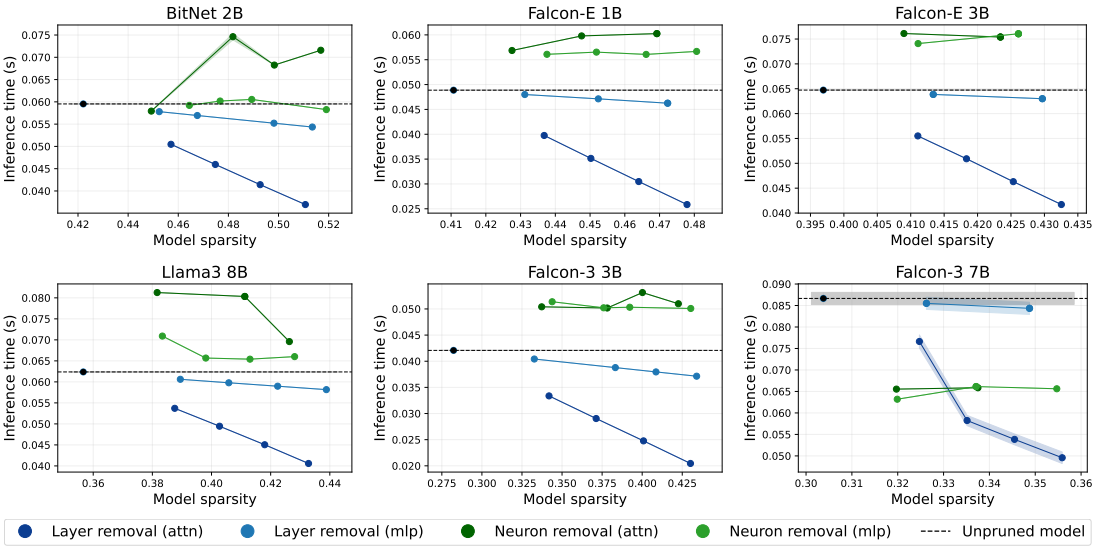


Figure 11: Inference time of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the inference time for the unpruned model. The black dot marks the sparsity of the original unpruned model. All shaded regions indicate variability across 128 samples (± standard deviation).

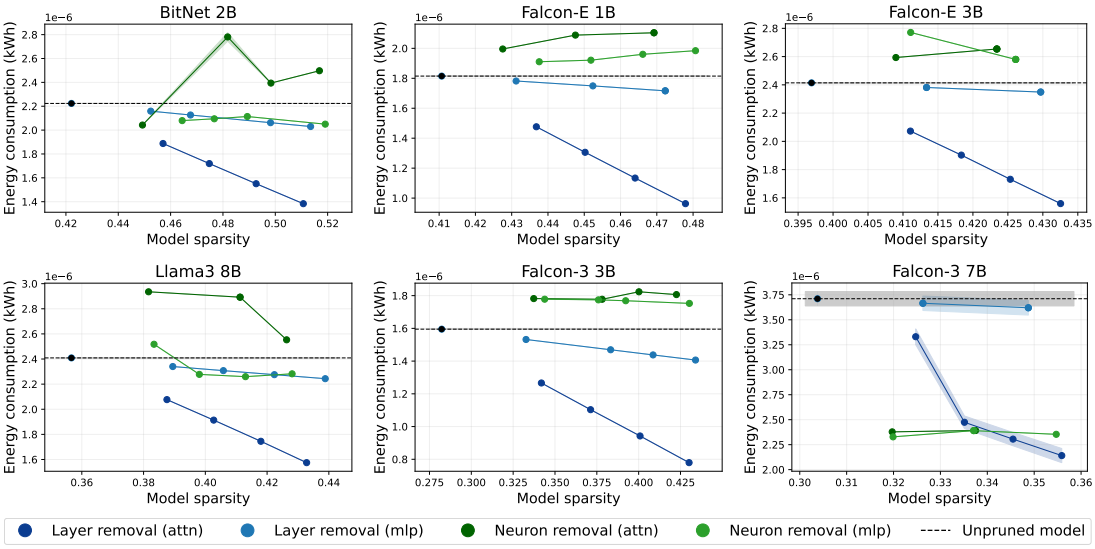


Figure 12: Energy consumption of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the energy consumption for the unpruned model. The black dot marks the sparsity of the original unpruned model. All shaded regions indicate variability across 128 samples (± standard deviation).

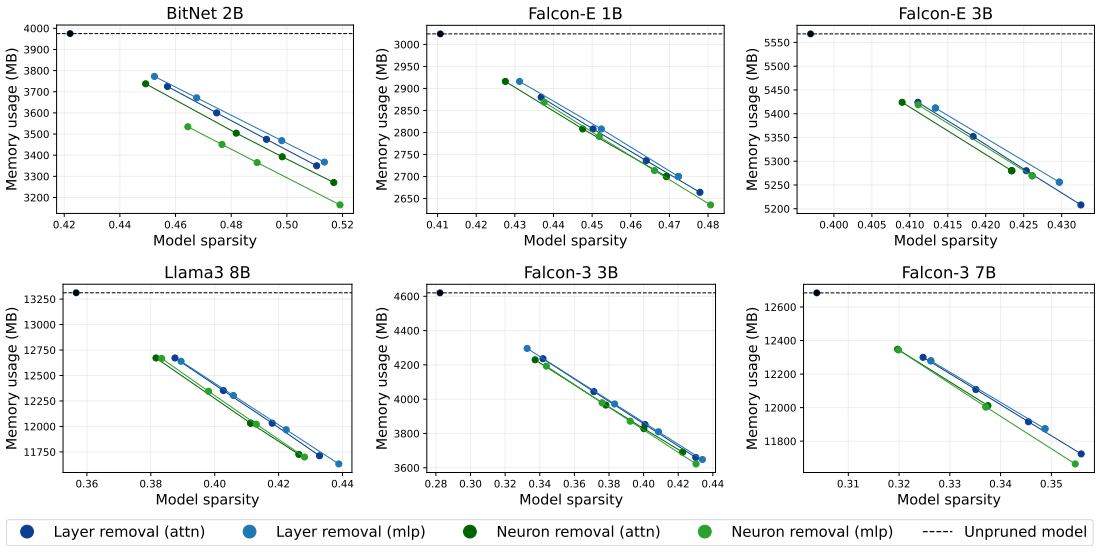


Figure 13: Memory requirements of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the memory requirements for the unpruned model. The black dot marks the sparsity of the original unpruned model.

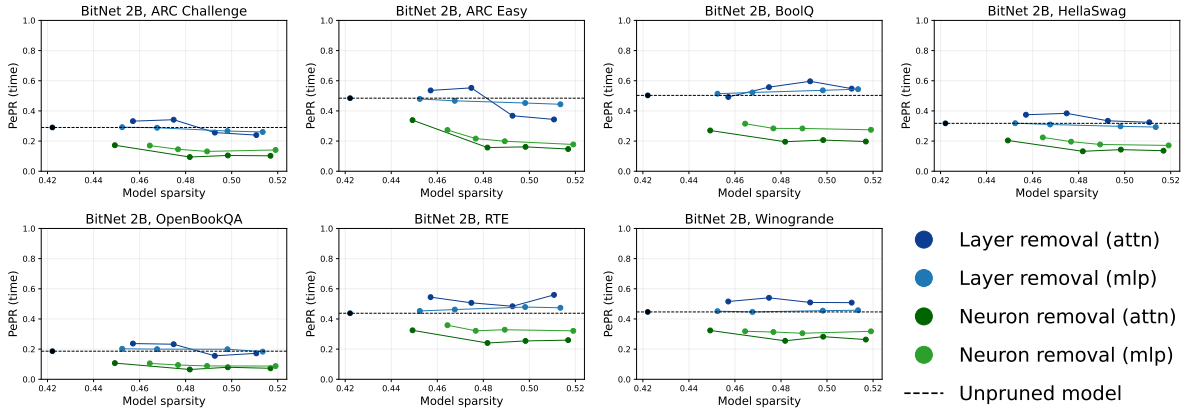


Figure 14: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

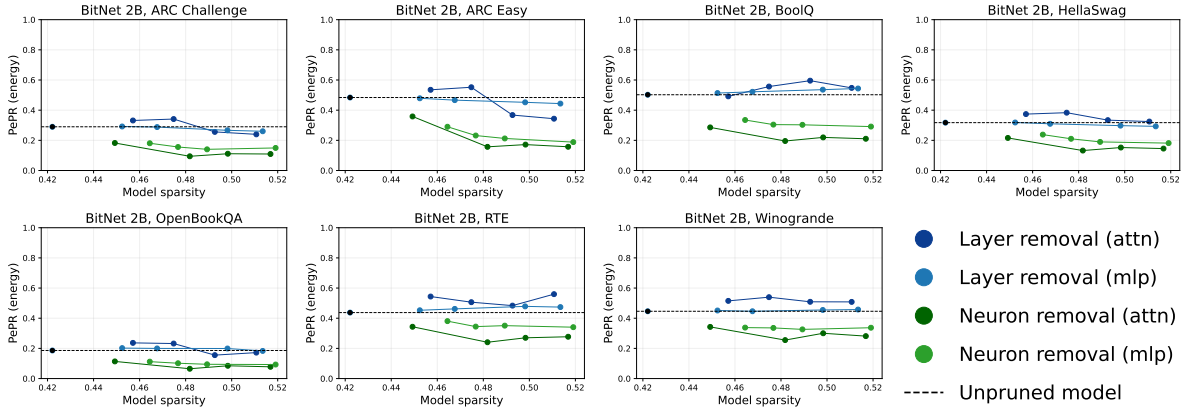


Figure 15: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

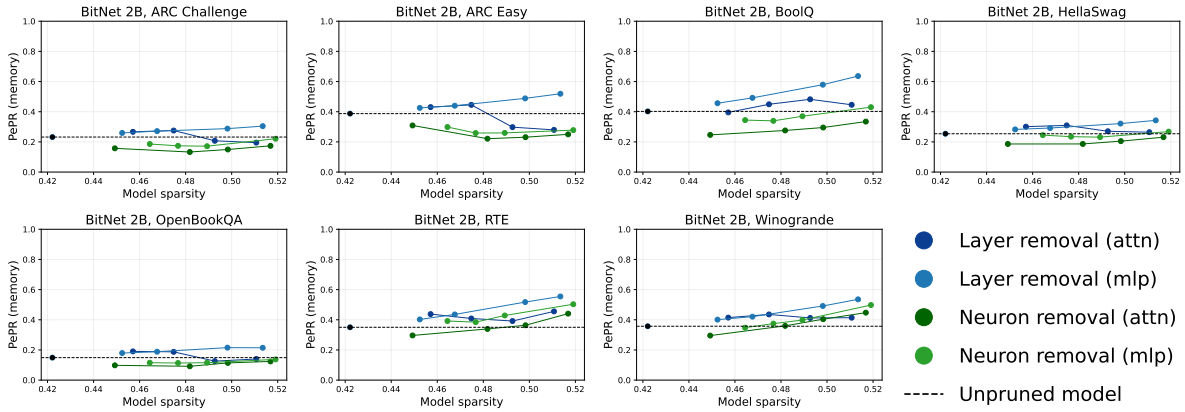


Figure 16: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

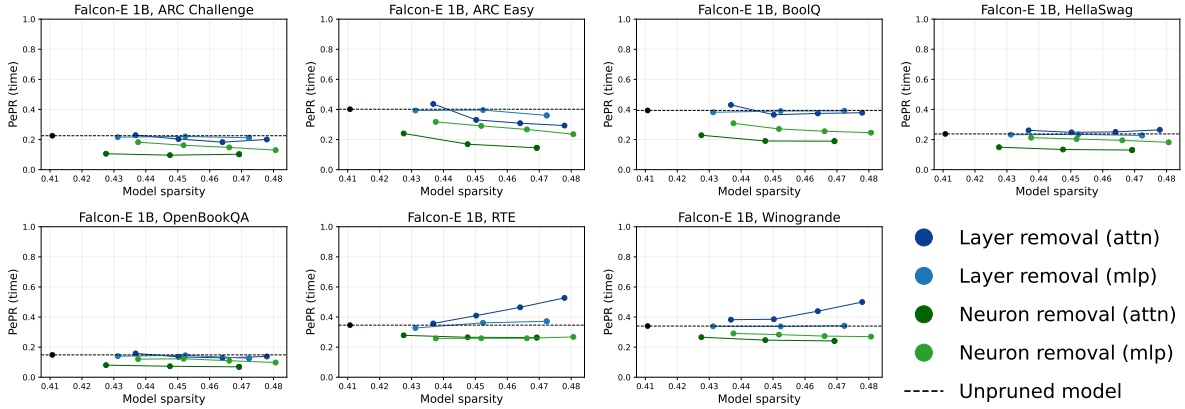


Figure 17: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

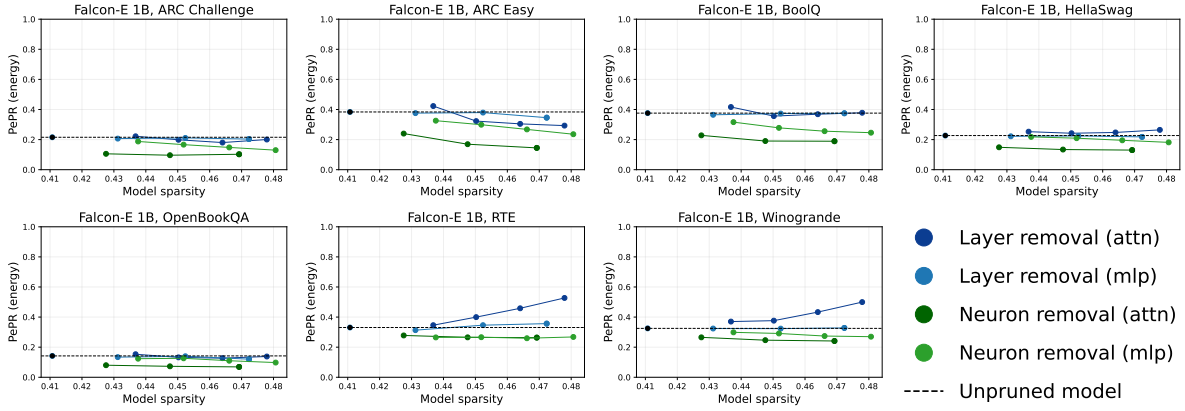


Figure 18: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

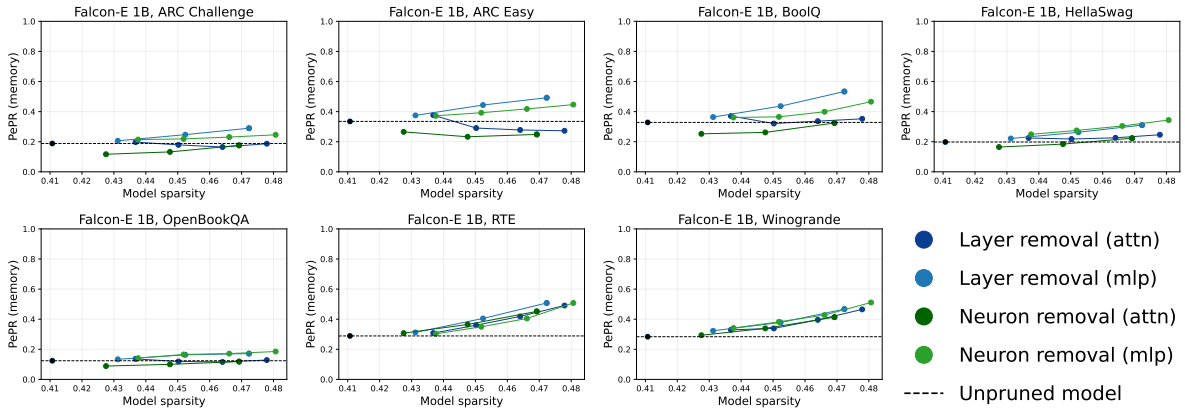


Figure 19: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

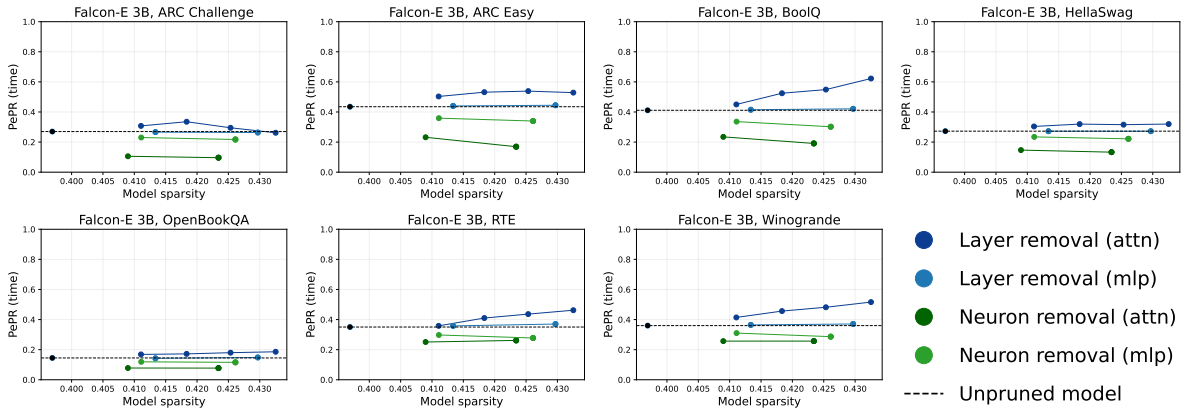


Figure 20: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

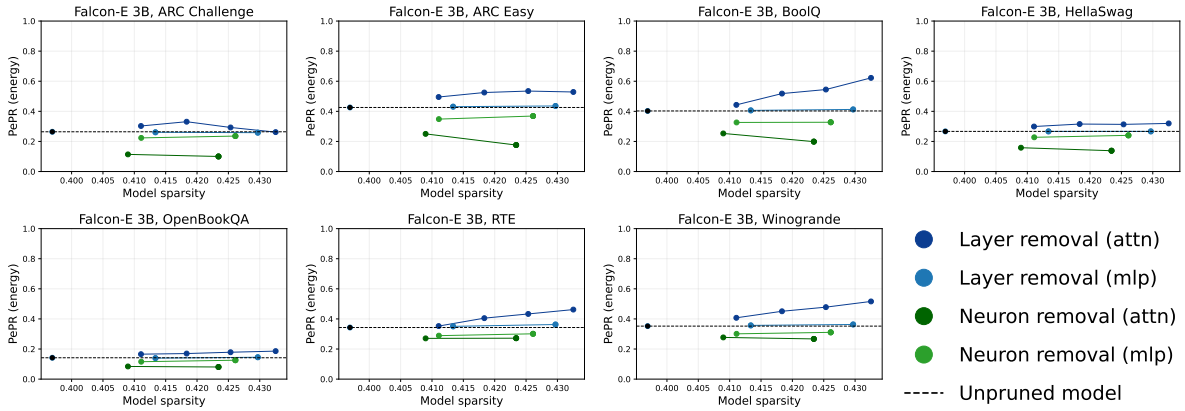


Figure 21: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

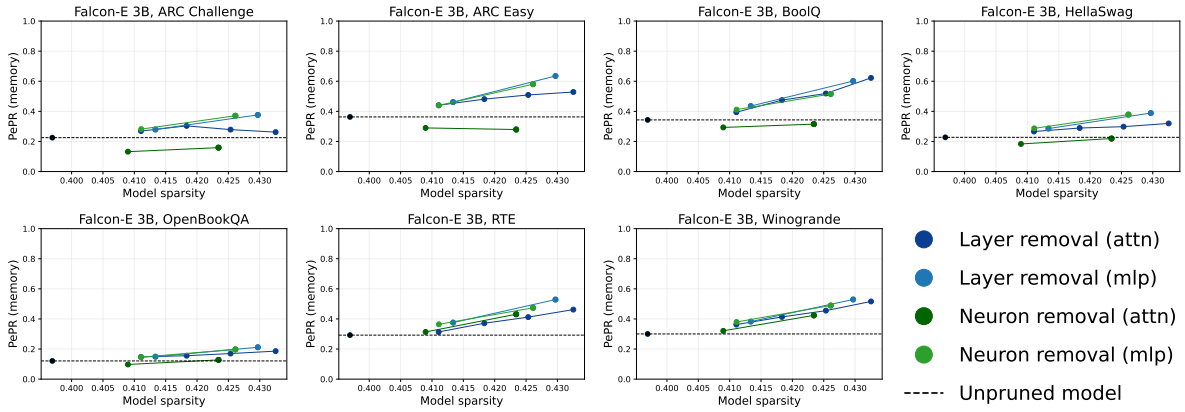


Figure 22: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

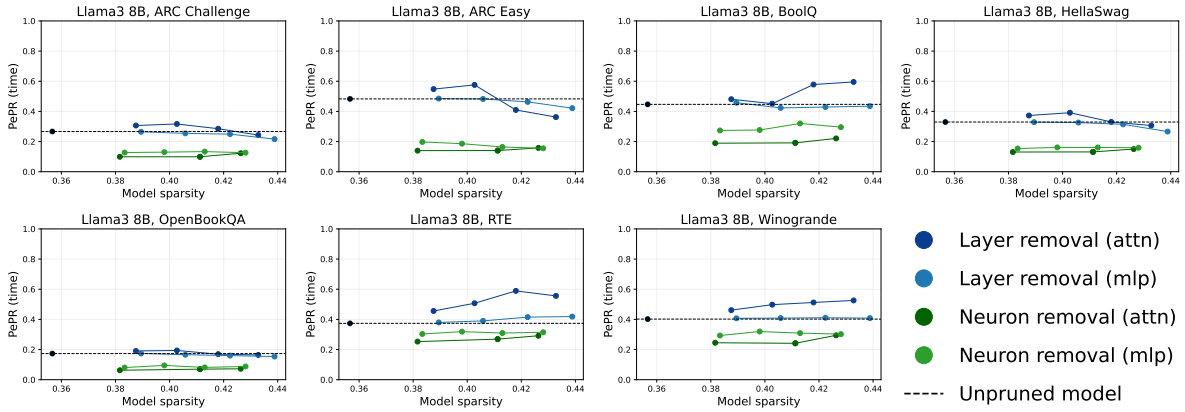


Figure 23: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

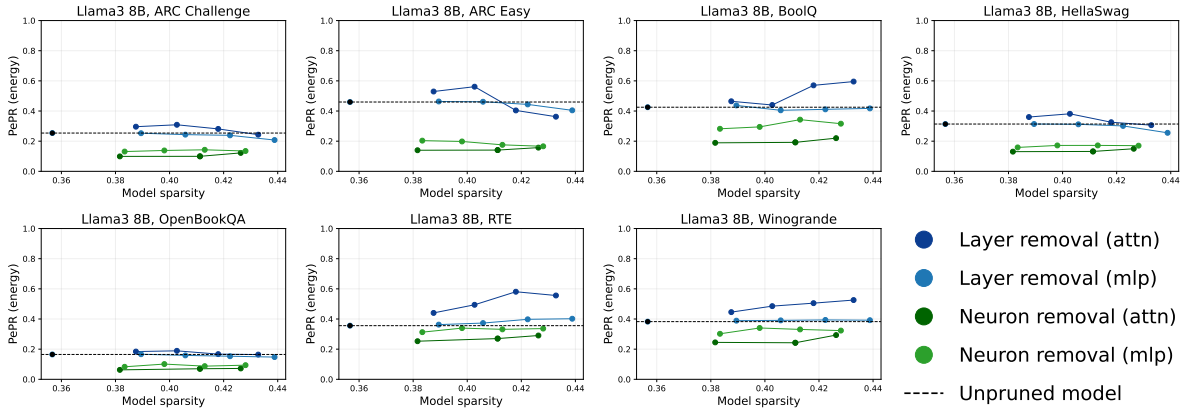


Figure 24: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

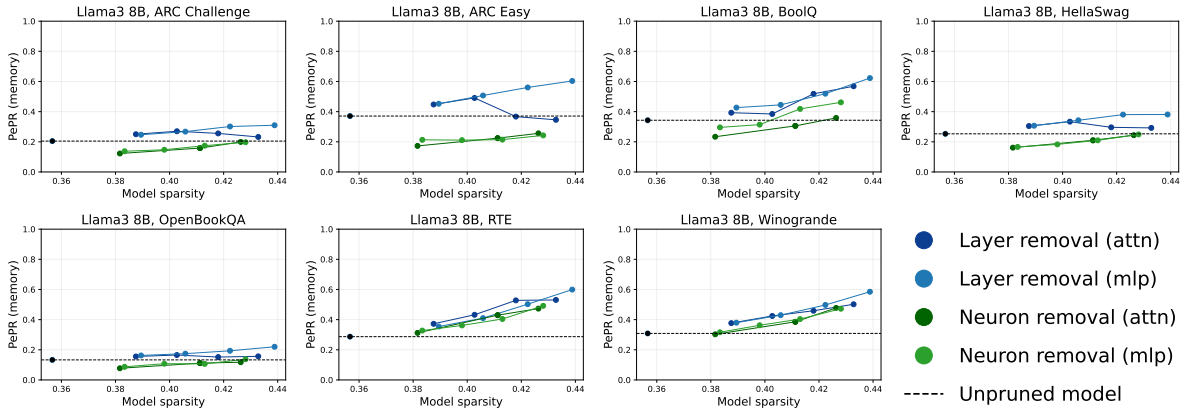


Figure 25: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

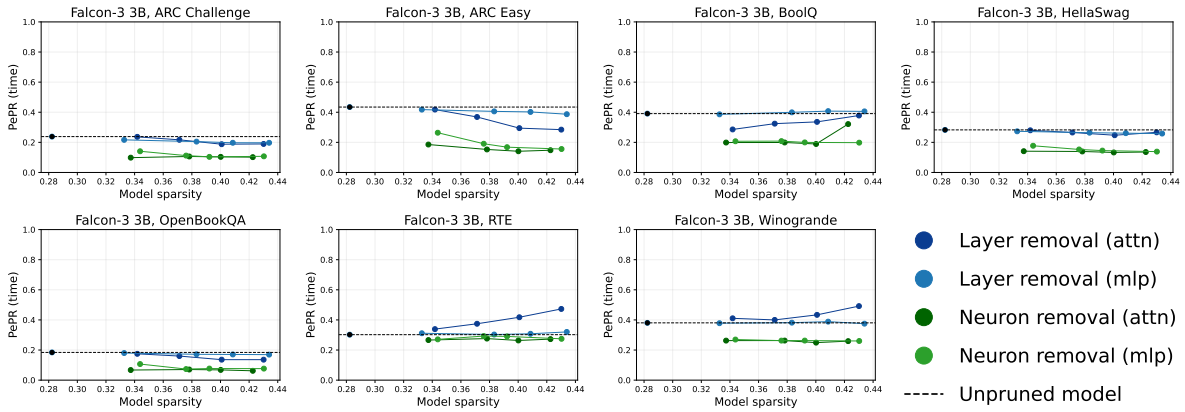


Figure 26: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

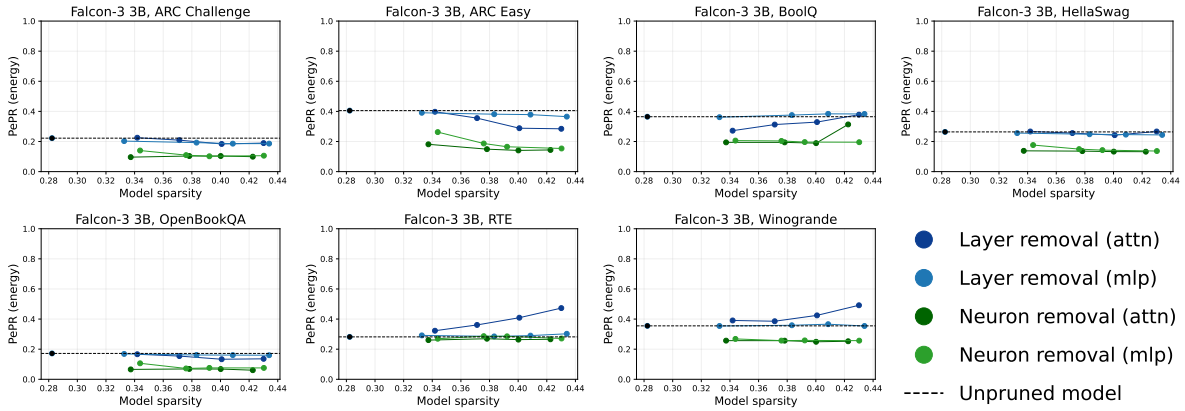


Figure 27: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

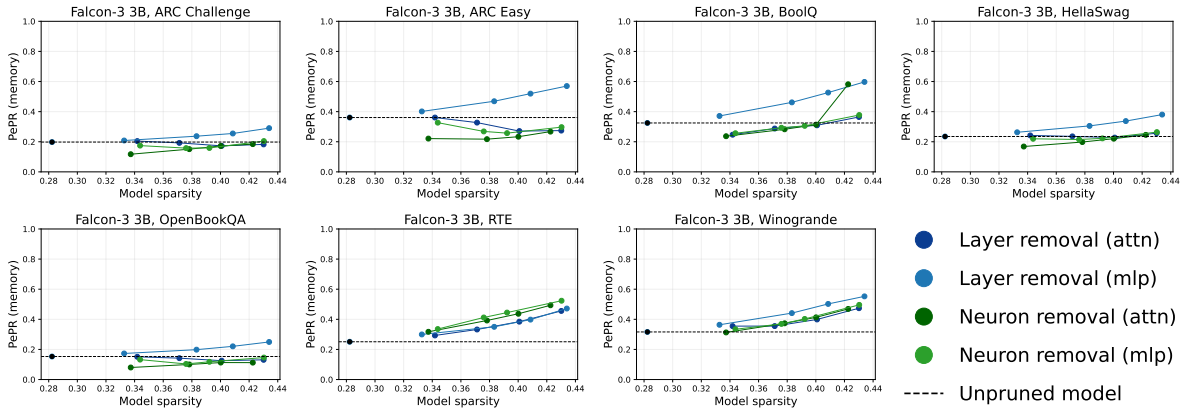


Figure 28: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

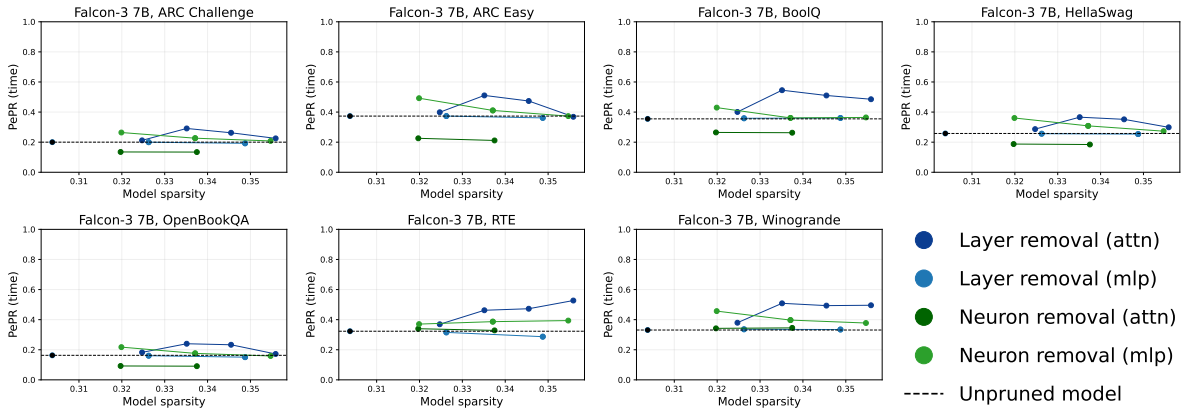


Figure 29: PePR time score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR time score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

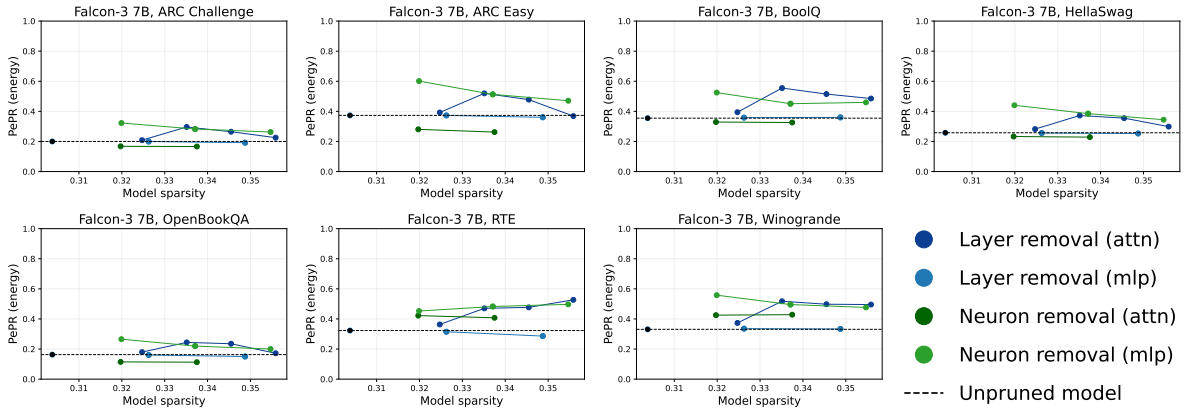


Figure 30: PePR energy score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR energy score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

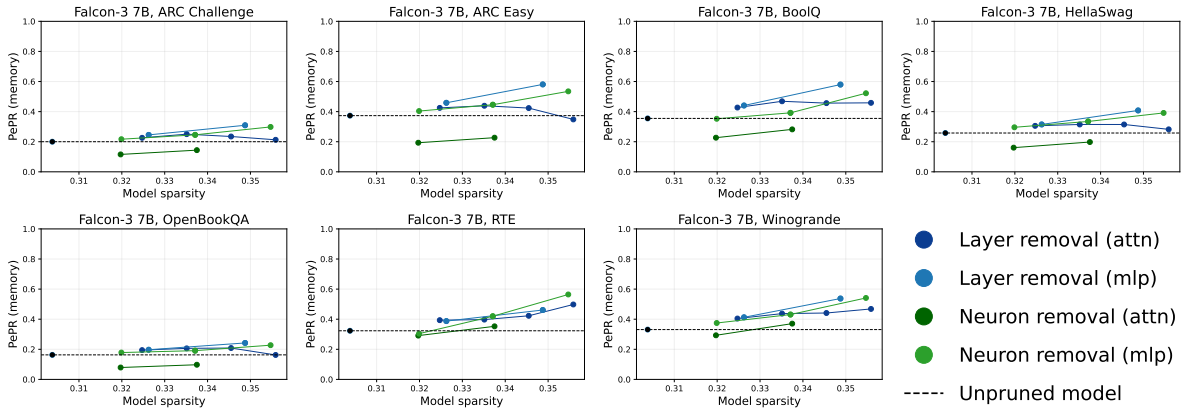


Figure 31: PePR memory score of pruned models as a function of sparsity (x-axis). Coloured lines represent different pruning methods, while the black dotted line indicates the PePR memory score for the unpruned model. The black dot marks the sparsity of the original unpruned model.

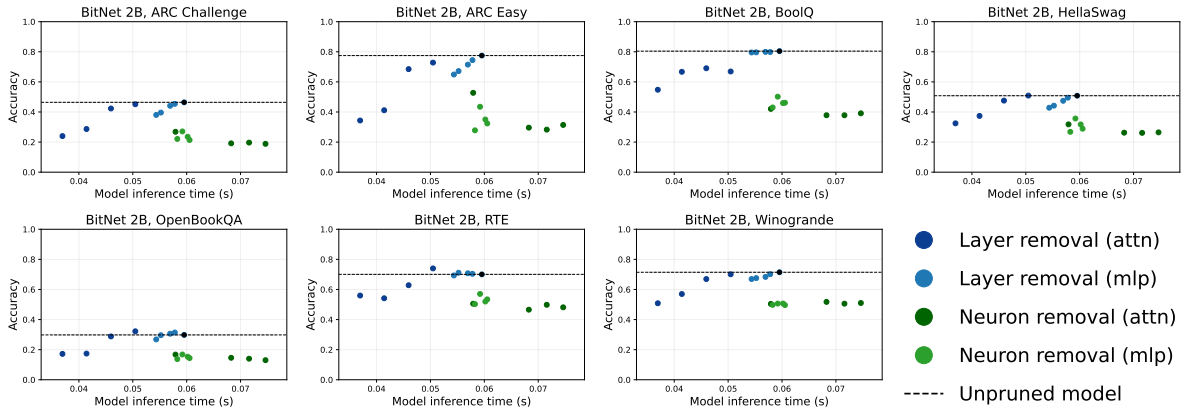


Figure 32: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

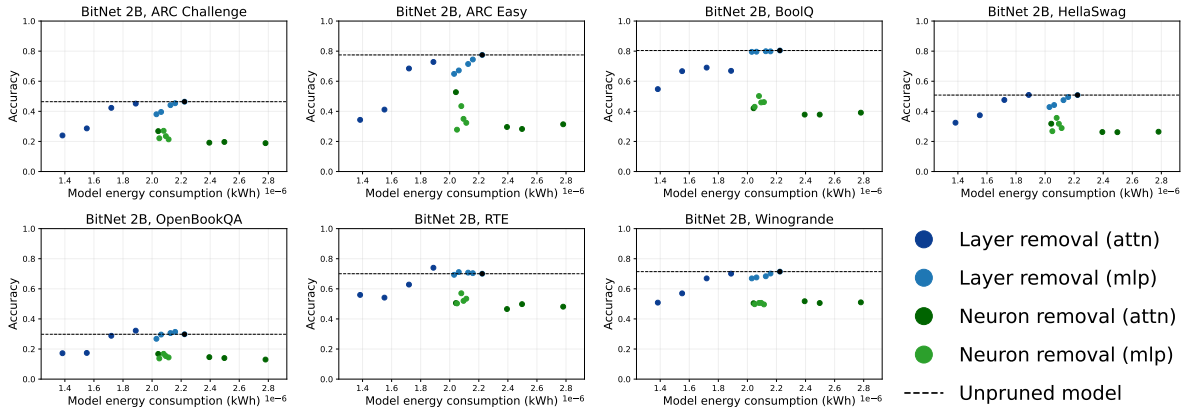


Figure 33: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

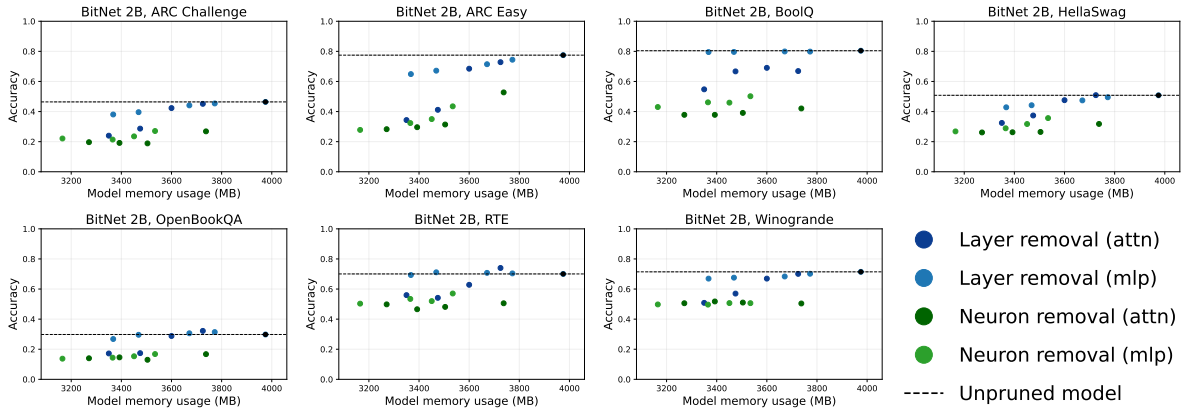


Figure 34: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

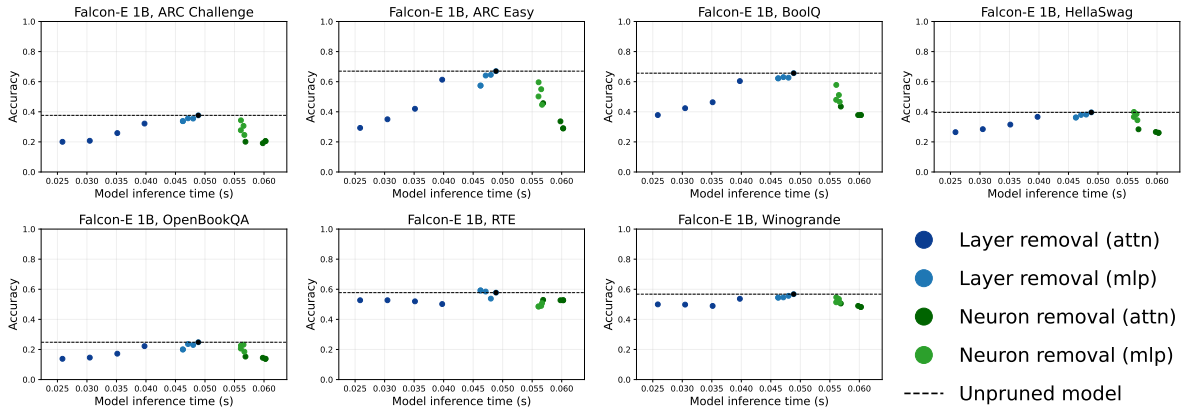


Figure 35: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

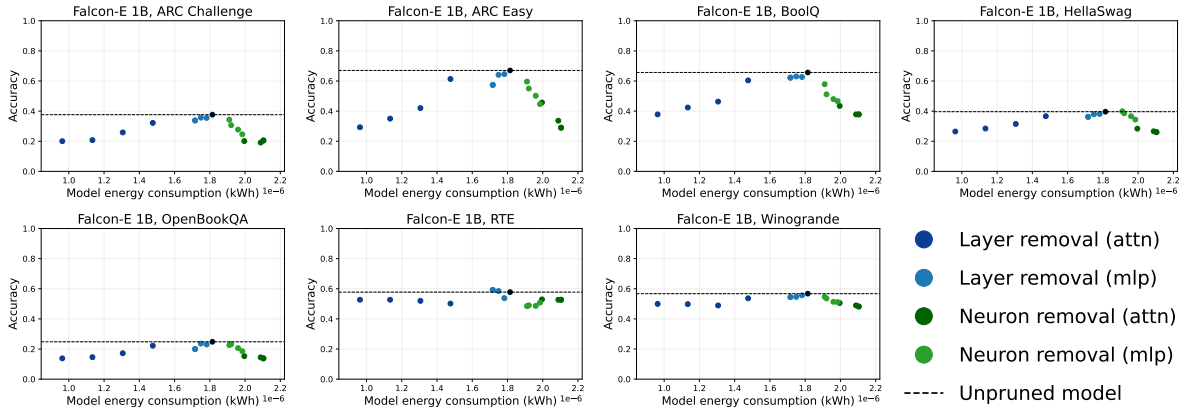


Figure 36: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

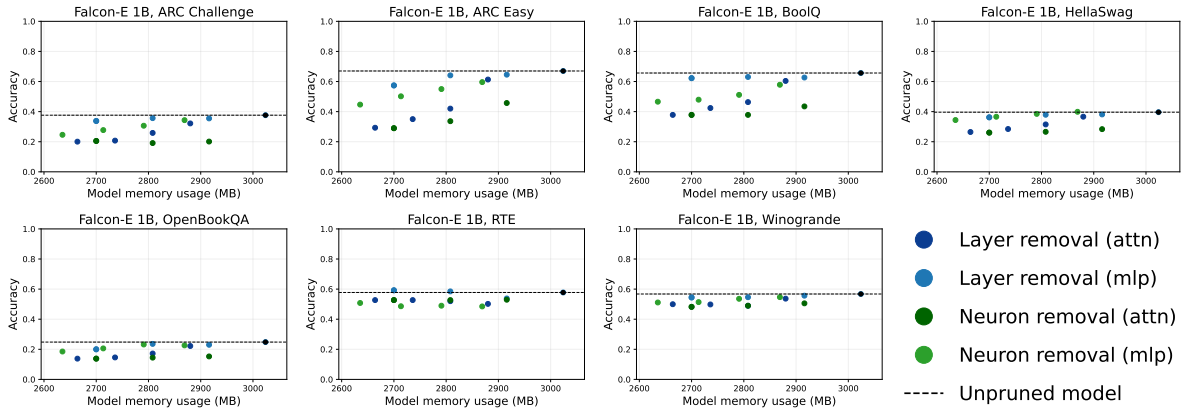


Figure 37: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

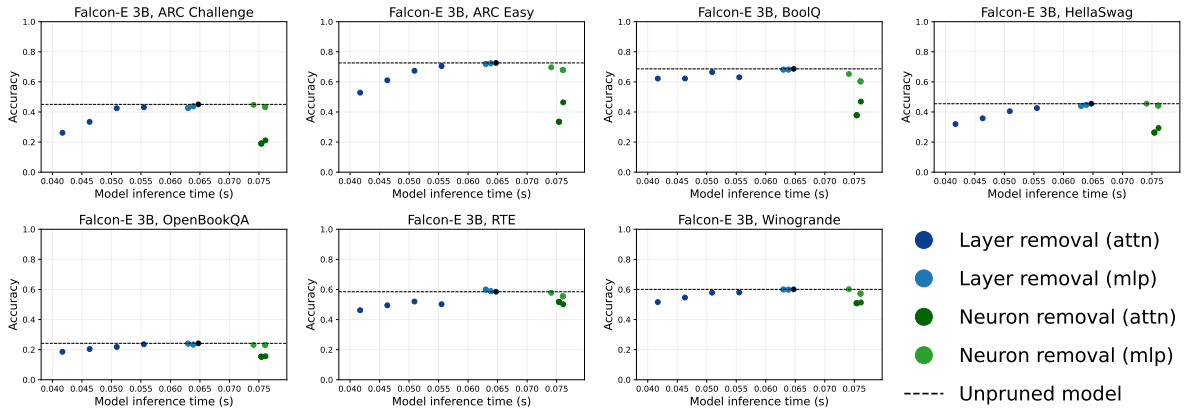


Figure 38: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

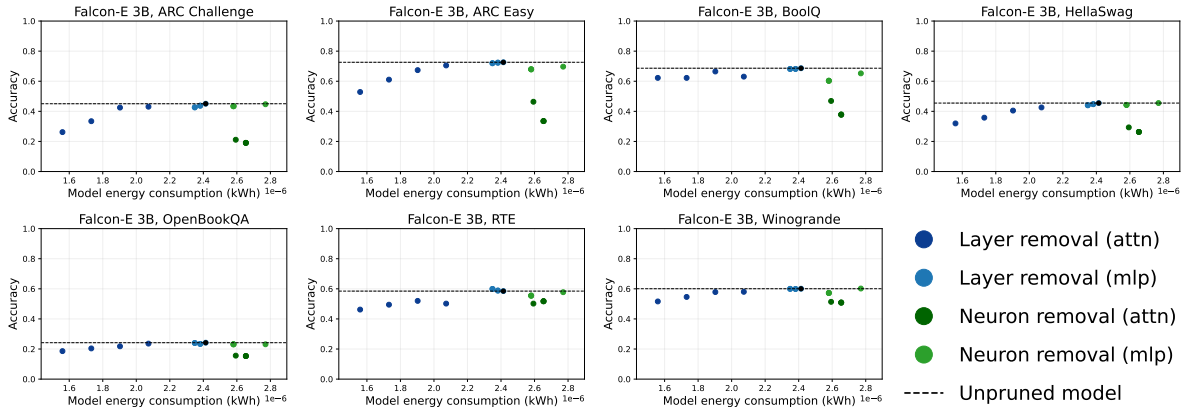


Figure 39: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

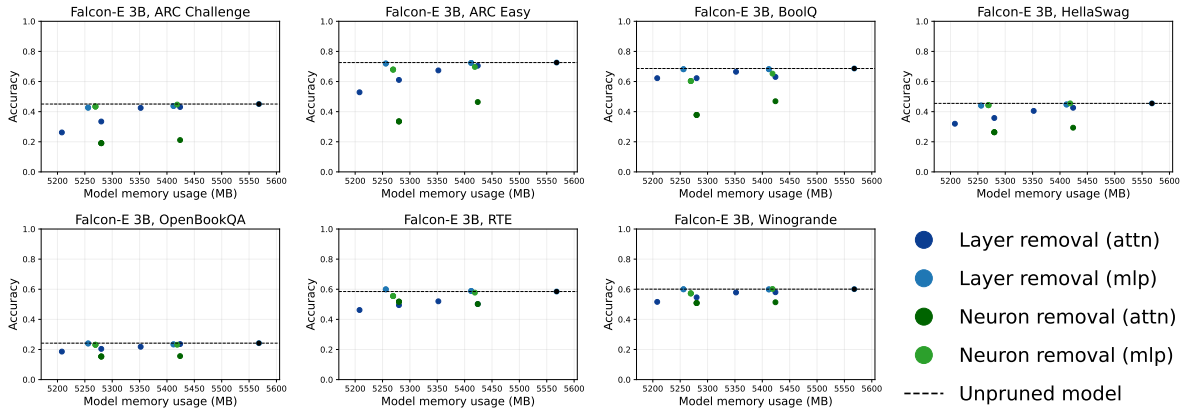


Figure 40: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

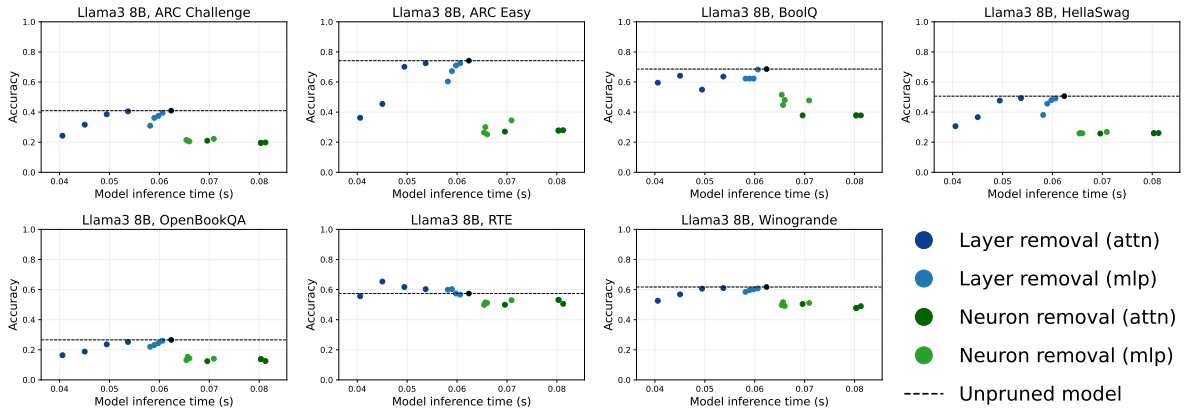


Figure 41: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

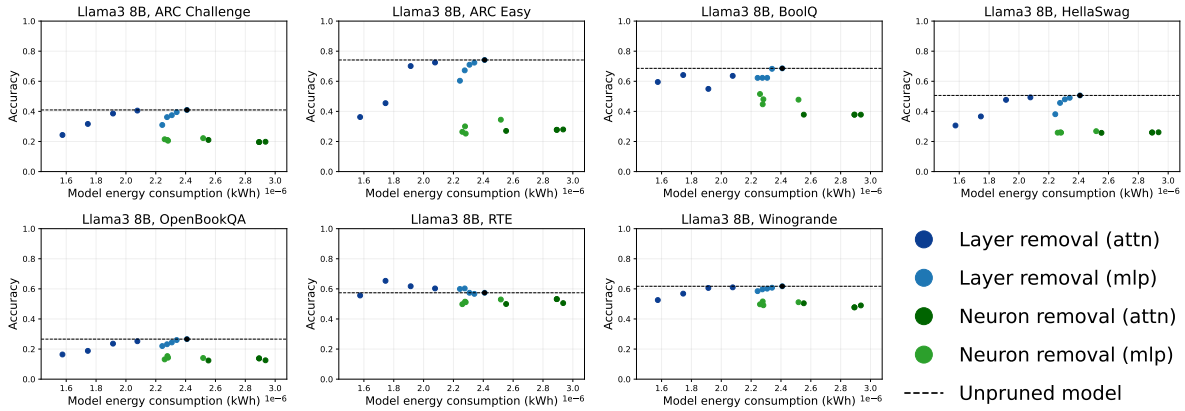


Figure 42: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

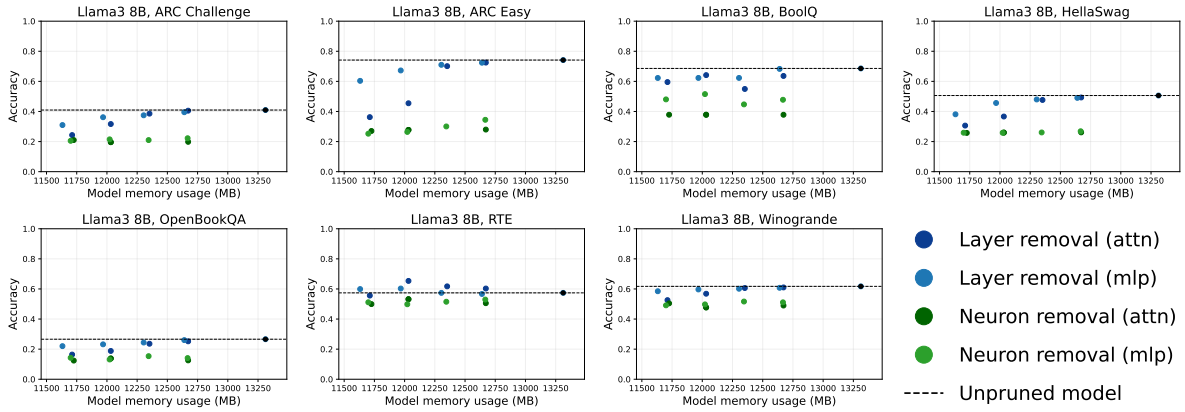


Figure 43: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

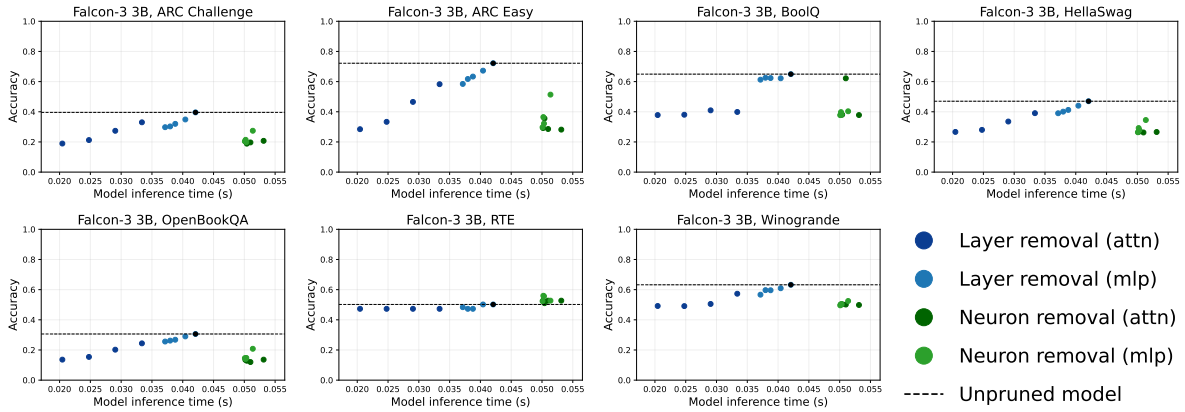


Figure 44: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

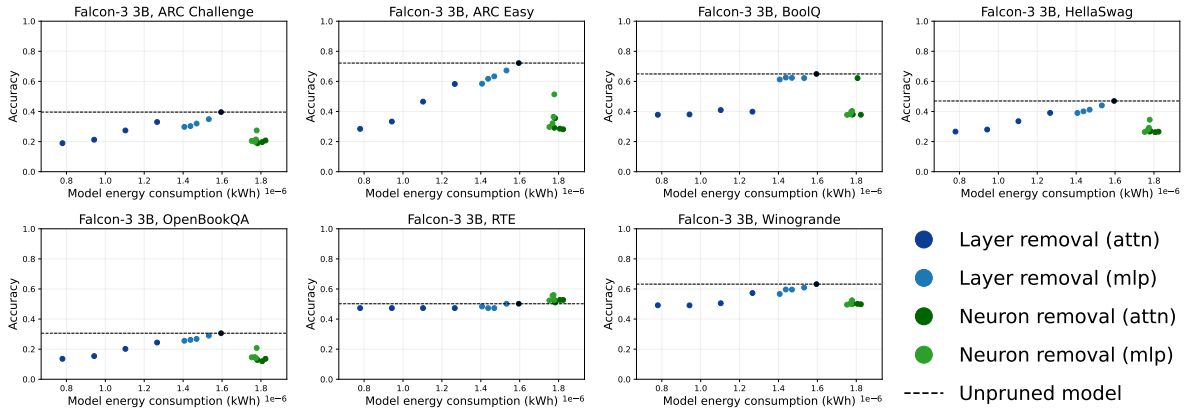


Figure 45: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

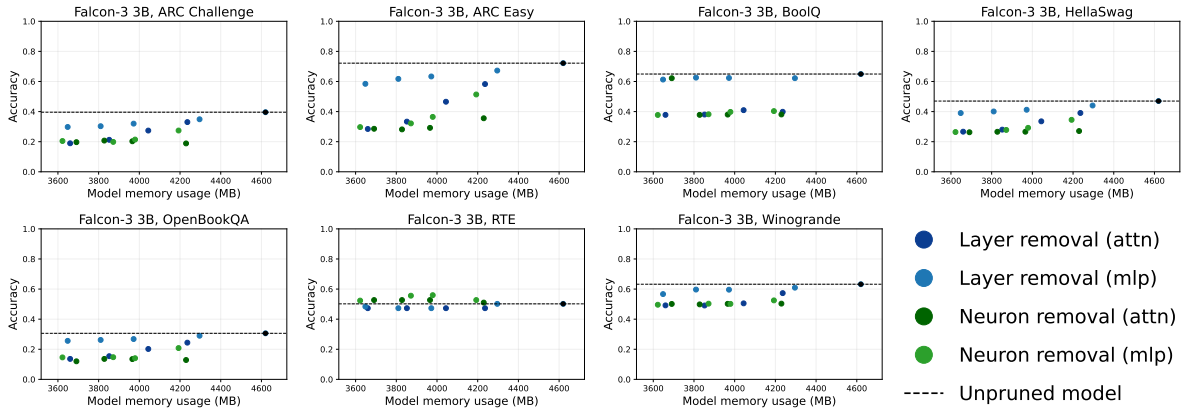


Figure 46: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

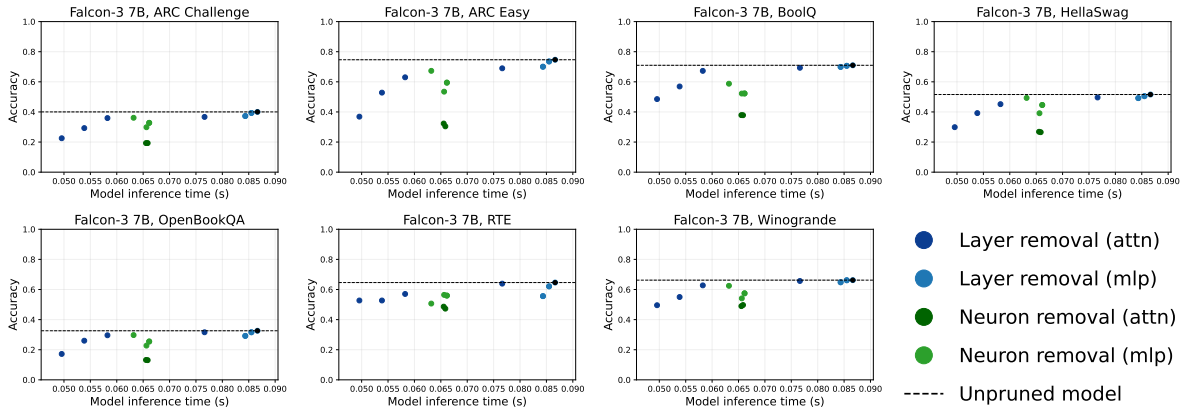


Figure 47: Inference time of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the inference time required by the unpruned model.

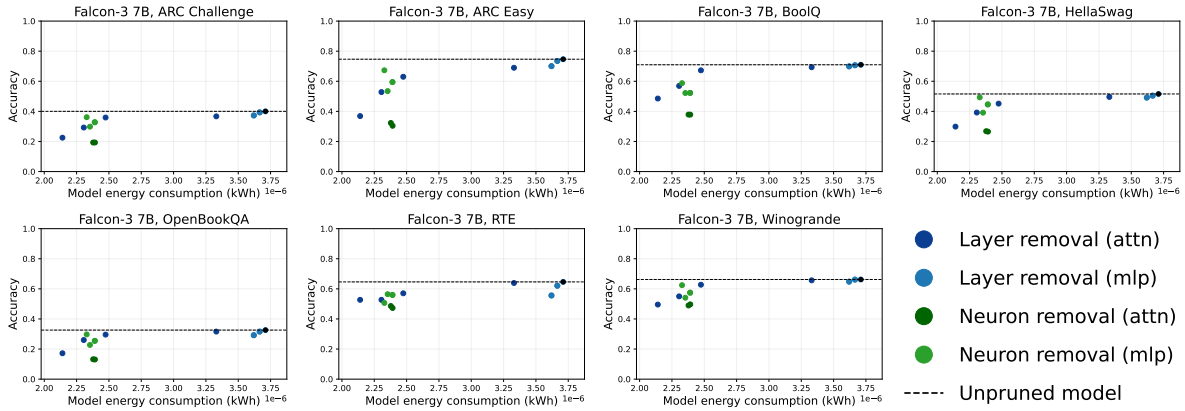


Figure 48: Energy consumption of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the energy consumption of the unpruned model.

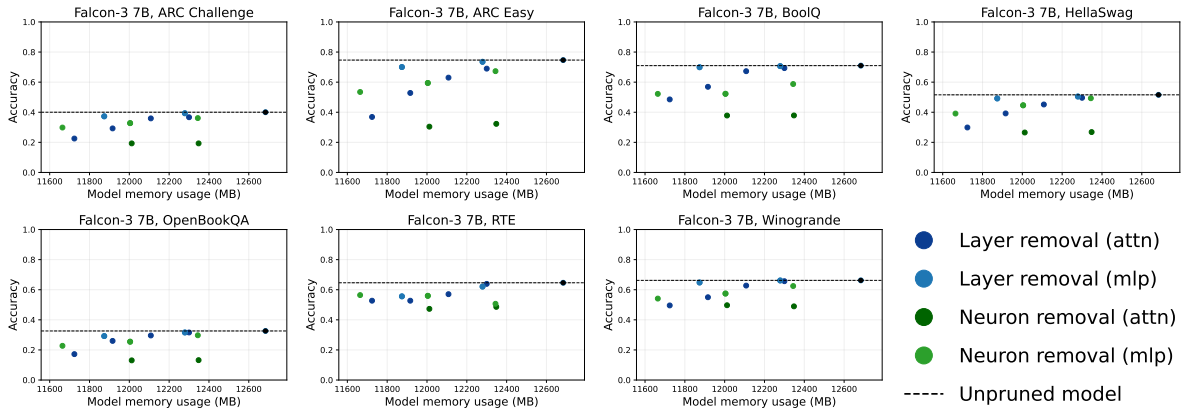


Figure 49: Memory requirements of pruned models against model accuracy (y-axis). Coloured dots represent different pruning methods, while the black line indicates the unpruned model accuracy. The black dot marks the memory requirements of the unpruned model.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.609	-	0.222	0.06	3975
Layer removal (attn)	8	0.589	0.02	0.189	0.05	3725
Layer removal (attn)	12	0.551	0.058	0.172	0.046	3600
Layer removal (attn)	16	0.432	0.177	0.155	0.041	3475
Layer removal (attn)	20	0.385	0.224	0.138	0.037	3350
Layer removal (mlp)	2	0.602	0.007	0.216	0.058	3773
Layer removal (mlp)	3	0.590	0.02	0.213	0.057	3671
Layer removal (mlp)	5	0.570	0.039	0.206	0.055	3469
Layer removal (mlp)	6	0.555	0.054	0.203	0.054	3368
Neuron removal (attn)	25	0.387	0.222	0.204	0.058	3738
Neuron removal (attn)	50	0.326	0.283	0.278	0.075	3504
Neuron removal (attn)	62	0.323	0.287	0.239	0.068	3393
Neuron removal (attn)	75	0.323	0.286	0.25	0.072	3271
Neuron removal (mlp)	12	0.401	0.208	0.208	0.059	3534
Neuron removal (mlp)	15	0.363	0.246	0.209	0.06	3451
Neuron removal (mlp)	18	0.352	0.257	0.211	0.061	3365
Neuron removal (mlp)	25	0.334	0.276	0.205	0.058	3165
Wanda	43	0.542	0.067	-	-	-
Wanda	46	0.497	0.112	-	-	-
Wanda	48	0.463	0.146	-	-	-
Wanda	50	0.430	0.179	-	-	-
Ria	43	0.446	0.163	-	-	-
Ria	46	0.386	0.223	-	-	-
Ria	48	0.362	0.247	-	-	-
Ria	50	0.351	0.258	-	-	-

Table 9: Aggregated BitNet 2B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.499	-	0.181	0.049	3024
Layer removal (attn)	8	0.452	0.047	0.148	0.04	2880
Layer removal (attn)	12	0.377	0.122	0.131	0.035	2808
Layer removal (attn)	16	0.348	0.151	0.113	0.03	2736
Layer removal (attn)	20	0.329	0.17	0.096	0.026	2664
Layer removal (mlp)	1	0.476	0.023	0.178	0.048	2916
Layer removal (mlp)	2	0.482	0.017	0.175	0.047	2808
Layer removal (mlp)	3	0.462	0.037	0.172	0.046	2700
Neuron removal (attn)	25	0.366	0.133	0.2	0.057	2916
Neuron removal (attn)	50	0.333	0.166	0.209	0.06	2808
Neuron removal (attn)	75	0.326	0.173	0.21	0.06	2700
Neuron removal (mlp)	6	0.454	0.045	0.191	0.056	2869
Neuron removal (mlp)	9	0.430	0.069	0.192	0.057	2791
Neuron removal (mlp)	12	0.404	0.095	0.196	0.056	2713
Neuron removal (mlp)	15	0.387	0.112	0.198	0.057	2635
Wanda	43	0.448	0.051	-	-	-
Wanda	46	0.412	0.087	-	-	-
Wanda	48	0.380	0.119	-	-	-
Ria	43	0.451	0.048	-	-	-
Ria	46	0.413	0.086	-	-	-
Ria	48	0.381	0.118	-	-	-

Table 10: Aggregated Falcon-E 1B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.535	-	0.241	0.065	5568
Layer removal (attn)	8	0.501	0.034	0.207	0.056	5424
Layer removal (attn)	12	0.498	0.037	0.19	0.051	5352
Layer removal (attn)	16	0.453	0.082	0.173	0.046	5280
Layer removal (attn)	20	0.414	0.121	0.156	0.042	5208
Layer removal (mlp)	1	0.530	0.005	0.238	0.064	5412
Layer removal (mlp)	2	0.530	0.005	0.235	0.063	5256
Neuron removal (attn)	25	0.373	0.162	0.259	0.076	5424
Neuron removal (attn)	50	0.335	0.2	0.265	0.075	5280
Neuron removal (mlp)	3	0.523	0.012	0.277	0.074	5419
Neuron removal (mlp)	6	0.503	0.032	0.258	0.076	5269
Wanda	37	0.518	0.017	-	-	-
Wanda	40	0.498	0.037	-	-	-
Wanda	41	0.490	0.045	-	-	-
Wanda	43	0.470	0.065	-	-	-
Ria	37	0.520	0.015	-	-	-
Ria	40	0.503	0.032	-	-	-
Ria	41	0.494	0.041	-	-	-
Ria	43	0.469	0.066	-	-	-

Table 11: Aggregated Falcon-E 3B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.543	-	0.241	0.062	13312
Layer removal (attn)	8	0.532	0.011	0.208	0.054	12672
Layer removal (attn)	12	0.510	0.033	0.191	0.049	12352
Layer removal (attn)	16	0.456	0.087	0.174	0.045	12032
Layer removal (attn)	20	0.393	0.15	0.157	0.041	11712
Layer removal (mlp)	2	0.532	0.01	0.234	0.061	12640
Layer removal (mlp)	3	0.515	0.028	0.231	0.06	12304
Layer removal (mlp)	4	0.506	0.036	0.228	0.059	11968
Layer removal (mlp)	5	0.474	0.068	0.224	0.058	11632
Neuron removal (attn)	25	0.320	0.223	0.294	0.081	12672
Neuron removal (attn)	50	0.323	0.22	0.289	0.08	12032
Neuron removal (attn)	62	0.321	0.222	0.255	0.07	11725
Neuron removal (mlp)	6	0.356	0.187	0.252	0.071	12668
Neuron removal (mlp)	9	0.343	0.2	0.228	0.066	12345
Neuron removal (mlp)	12	0.340	0.203	0.226	0.065	12022
Neuron removal (mlp)	15	0.335	0.208	0.228	0.066	11700
Wanda	37	0.493	0.05	-	-	-
Wanda	40	0.456	0.087	-	-	-
Wanda	43	0.398	0.145	-	-	-
Ria	37	0.489	0.054	-	-	-
Ria	40	0.449	0.094	-	-	-
Ria	43	0.390	0.153	-	-	-

Table 12: Aggregated Llama3 8B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.525	-	0.16	0.042	4620
Layer removal (attn)	8	0.428	0.098	0.127	0.033	4236
Layer removal (attn)	12	0.381	0.145	0.11	0.029	4044
Layer removal (attn)	16	0.332	0.193	0.094	0.025	3852
Layer removal (attn)	20	0.317	0.208	0.078	0.02	3660
Layer removal (mlp)	2	0.498	0.027	0.153	0.04	4296
Layer removal (mlp)	4	0.475	0.05	0.147	0.039	3972
Layer removal (mlp)	5	0.468	0.057	0.144	0.038	3810
Layer removal (mlp)	6	0.456	0.069	0.141	0.037	3648
Neuron removal (attn)	37	0.334	0.191	0.178	0.05	4230
Neuron removal (attn)	62	0.329	0.196	0.178	0.05	3966
Neuron removal (attn)	75	0.328	0.198	0.182	0.053	3828
Neuron removal (attn)	88	0.359	0.166	0.181	0.051	3691
Neuron removal (mlp)	12	0.399	0.126	0.178	0.051	4193
Neuron removal (mlp)	18	0.353	0.172	0.177	0.05	3979
Neuron removal (mlp)	21	0.341	0.184	0.177	0.05	3872
Neuron removal (mlp)	28	0.330	0.196	0.175	0.05	3622
Wanda	34	0.490	0.035	-	-	-
Wanda	37	0.460	0.065	-	-	-
Wanda	40	0.439	0.086	-	-	-
Wanda	43	0.409	0.116	-	-	-
Ria	34	0.491	0.034	-	-	-
Ria	37	0.473	0.052	-	-	-
Ria	40	0.443	0.082	-	-	-
Ria	43	0.408	0.117	-	-	-

Table 13: Aggregated Falcon-3 3B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Pruning Algorithm	Threshold	Accuracy	Accuracy Drop	Energy	Time	Memory
Unpruned	-	0.572	-	0.371	0.087	12684
Layer removal (attn)	8	0.551	0.021	0.333	0.077	12300
Layer removal (attn)	12	0.515	0.057	0.247	0.058	12108
Layer removal (attn)	16	0.446	0.127	0.231	0.054	11916
Layer removal (attn)	20	0.368	0.205	0.214	0.05	11724
Layer removal (mlp)	1	0.563	0.01	0.367	0.086	12279
Layer removal (mlp)	2	0.537	0.035	0.362	0.084	11874
Neuron removal (attn)	25	0.324	0.248	0.238	0.066	12348
Neuron removal (attn)	50	0.320	0.252	0.239	0.066	12012
Neuron removal (mlp)	3	0.506	0.066	0.233	0.063	12344
Neuron removal (mlp)	6	0.469	0.104	0.239	0.066	12004
Neuron removal (mlp)	9	0.440	0.132	0.235	0.066	11664
Wanda	31	0.563	0.01	-	-	-
Wanda	34	0.547	0.026	-	-	-
Wanda	35	0.533	0.039	-	-	-
Ria	31	0.565	0.008	-	-	-
Ria	34	0.544	0.028	-	-	-
Ria	35	0.534	0.039	-	-	-

Table 14: Aggregated Falcon-3 7B results for each pruning configuration. Results are averaged across datasets and, for pruned configurations, across random seeds. No information is reported for Energy, Time, and Memory for Wanda and Ria as unstructured pruning algorithms do not lead to direct reductions in those measures. Energy is reported in Kwh and exponentiated by 1e5 to appreciate changes. Time is reported in seconds and memory is reported in MB.

Does Tropical Adaptation Help Plant Language Models Generalize? A Split-Sensitivity Evaluation

Gabriela Morales Soto
Institute of Computing - UNICAMP
Campinas-SP, Brazil, 13083-852
g299213@ic.unicamp.br

Helio Pedrini
Institute of Computing - UNICAMP
Campinas-SP, Brazil, 13083-852
helio@ic.unicamp.br

Abstract

Plant species assemblages can contain climate-related information, but it is unclear whether plant language models remain useful when observations are incomplete. We compare a sparse LightGBM baseline with Pl@ntBERT and a tropically adapted variant combining vocabulary expansion and domain-adaptive pretraining. The models predict four WorldClim variables from species composition in São Paulo state under progressively shortened species lists. The clearest pattern occurs for precipitation of the driest month (BIO14). On the original split, A2-MT shows positive differences over LightGBM under severe incompleteness after hierarchical bootstrap and multiplicity correction. Because BIO14 was selected post hoc, we retrain the models on five alternative assignments of a nominal 40 km spatial-block design. LightGBM is higher on complete lists in four of five partitions, while positive BIO14 differences under incomplete lists are directionally more stable but retain corrected support in only two partitions. Additional checks show that the pattern persists with a longer context window, weakens under frequency-weighted removal, and narrows to Max-10 and Max-5 when both model families receive truncation-augmented training. These results show that within-split robustness does not substitute for retraining across alternative spatial partitions.

1 Introduction

A species assemblage, the set of species observed in a locality, can carry information about climate because species differ in their environmental tolerances. This principle underlies methods such as the Modern Analogue Technique (Imbrie and Kipp, 1971; Overpeck et al., 1985), which relates assemblage composition to environmental conditions observed in reference data.

Biodiversity assemblage data resemble other incomplete environmental records used elsewhere in this workshop: both require inferring a climate-related signal from partial, unevenly sampled observations. We frame our contribution as methodological transfer of NLP-style sequence models to non-textual structured records under spatial dependence, with climate inference as the motivating application rather than a demonstrated operational tool.

We study whether a plant language model can use this signal in a tropical setting. Pl@ntBERT (Leblanc et al., 2024, 2025) applies a BERT-style sequence model, an architecture originally built to learn from sentences by predicting words from their context, to species lists instead of text. Its pretraining data, however, are predominantly European vegetation plots, so many tropical taxa and co-occurrence patterns may be underrepresented in what it originally learned.

We evaluate four WorldClim bioclimatic targets: annual mean temperature (BIO1), annual precipitation (BIO12), precipitation of the driest month (BIO14), and precipitation seasonality (BIO15). BIO14 showed the strongest pattern on the original split and was therefore selected post hoc for the retraining-based split-sensitivity analysis. We treat this follow-up as an internal sensitivity analysis rather than a confirmatory test.

The practical motivation is incomplete biodiversity observation. Species occurrence databases such as the Global Biodiversity Information Facility (GBIF) can have uneven sampling across space and taxa (Hughes et al., 2021). In our corpus, observed richness is strongly right-skewed: the median 5 km grid cell contains 24 species, compared with a mean of 58.9. We therefore ask whether a language-model pipeline retains useful climate signal when species lists are shortened, and whether it outperforms a sparse LightGBM baseline built from binary species-presence features under those conditions.

A broader methodological question is whether

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 85–98
October 28, 2026 ©2026 Association for Computational Linguistics

robustness within one fixed split implies robustness to the choice of split itself. We provide three main contributions: (i) we evaluate how a combined tropical-adaptation pipeline—train-only vocabulary expansion followed by domain-adaptive pretraining—changes Pl@ntBERT transfer across continuous climate regression and an exploratory phytophysiognomy-classification task; (ii) we compare the adapted model with a sparse LightGBM baseline under simulated species-list incompleteness; and (iii) for the strongest post-hoc target, BIO14, we distinguish within-split resampling from full retraining across five alternative assignments of one nominal 40 km spatial-block design.

2 Related Work

Assemblage composition has long been used to infer environmental conditions. The Modern Analogue Technique compares target assemblages with modern reference assemblages observed under known climates (Imbrie and Kipp, 1971; Overpeck et al., 1985), while the coexistence approach reconstructs climatic intervals from overlapping taxon tolerances (Mosbrugger and Utescher, 1997). Our study examines a related inference problem using large occurrence datasets and a domain-adapted plant language model.

Pl@ntBERT (Leblanc et al., 2024, 2025) adapts a BERT-style architecture (Devlin et al., 2019) to abundance-ordered species lists and evaluates tasks including masked-species prediction and habitat classification. Our study examines transfer to tropical assemblages and continuous bioclimatic regression. Domain-adaptive pretraining (DAPT), continued pretraining on in-domain unlabeled data before task fine-tuning, has improved downstream performance in other domains (Gururangan et al., 2020). We test whether vocabulary expansion combined with DAPT improves performance in this setting and whether any observed gain persists across data splits.

Spatially structured evaluation is important when nearby observations share environmental information. Roberts et al. (2017) show that conventional cross-validation can underestimate predictive error under spatial dependence and recommend structure-aware splitting. Ploton et al. (2020) similarly report weaker ecological prediction under spatial validation. These results motivate our distinction between resampling an existing test set and retraining on a newly constructed spatial partition.

3 Data and Experimental Design

This section describes the data, model variants, and experimental protocols used in our evaluation.

3.1 Data Sources and Analysis Subsets

We use three data sources with distinct roles. GBIF occurrence records provide the plant observations used to construct cell-level species assemblages and binary species-presence features. IF-SP 2020 vegetation polygons (Instituto Florestal, 2020) define the analysis subset and provide labels for the exploratory classification task reported in Appendix E. WorldClim 2.1 (Fick and Hijmans, 2017) provides the four regression targets and is never used as model input.

For GBIF (GBIF.org, 2026), we retain plant records with valid coordinates, species-level identification, and `occurrenceStatus == PRESENT`, after applying the geolocation filters used in preprocessing. Records are assigned to a 5 km grid, and cells with fewer than 10 records or 5 species are removed. This yields an ecological corpus of 4,014 cells, 503,859 occurrence records, and 13,355 unique species.

To define the analysis subset, we intersect grid cells with IF-SP 2020 vegetation polygons and retain cells whose dominant mapped class covers at least 30% of the intersected IF-SP area. The resulting IFSP30 subset contains 1,077 cells. A post hoc geometric check found that 10.5% of these cells (113 of 1,077) have centroids falling just outside São Paulo's official political boundary, at a maximum distance of 3.5 km (median 950 m); this is consistent with a boundary mismatch between the IF-SP 2020 polygons and the administrative boundary. A border-cell exclusion and matched placebo analysis suggest that these cells do not uniquely drive the reported BIO14 pattern (Appendix F). The original fixed partition contains 751 training and 326 test cells; it was inherited from an earlier phase of the project, and its exact generation procedure (random, spatial, or otherwise) could not be reconstructed from the retained artifacts. This uncertainty is one motivation for the retraining-based split-sensitivity analysis below. After excluding cells with missing climate targets, the regression analysis uses 749 training and 321 test cells. Each cell is represented as a deduplicated sequence of species names in case-insensitive lexicographic order. This representation differs from the abundance-oriented ordering used in Pl@ntBERT pretraining,

a mismatch considered in the Limitations.

3.2 Bioclimatic Targets

The four targets defined in Section 1 are extracted from WorldClim 2.1 (Fick and Hijmans, 2017) at 2.5 arc-min resolution by cell centroid. At São Paulo's latitude, this corresponds to approximately 4.25 to 4.64 km, comparable to the 5 km analysis grid. Their means (SDs) in the analyzed data are BIO1 19.6°C (2.3°C), BIO12 1628 mm (407 mm), BIO14 49 mm (26 mm), and BIO15 55.7 (12.9). Climate variables are used only as regression targets and are never provided as model inputs.

3.3 Model Variants

Throughout the paper, A0 and A2 denote the two adaptation states evaluated in this study. A0 refers to direct transfer from the original Pl@ntBERT checkpoint without tropical adaptation. A2 refers to the tropically adapted model obtained through the expansion of the train-only vocabulary followed by domain-adaptive pretraining in the training cells. The suffixes follow the experimental configurations used here: FT identifies the direct full-fine-tuning configuration used with A0, whereas MT and ST distinguish multi-target and single-target fine-tuning of A2. The resulting configurations are A0-FT, A2-MT, and A2-ST.

Our baseline consists of four independent LightGBM regressors, one per climate target, trained on binary species-presence features. The feature space is restricted to species observed in the training set. Exact hyperparameters are reported in Appendix D.

A0-FT starts from the original Pl@ntBERT checkpoint (Leblanc et al., 2024, 2025) and fine-tunes the full model end to end with a linear regression head that jointly predicts the four climate targets. No tropical adaptation is applied.

A2-MT starts from the A2 model defined above and is fine-tuned end to end to jointly predict the same four climate targets. We repeat this procedure across five training seeds, 42 through 46. Exact training settings are reported in Appendix D.

A2-ST uses the same A2 model but is fine-tuned only for precipitation of the driest month. It serves as a secondary control for whether the pattern observed for this variable depends on joint supervision from the four climate targets. A2-ST and the frozen-probe analyses are evaluated only on the original fixed split and are not retrained on the new spatial-block split.

We also freeze the A0 and A2 representations and fit ridge regressors under three pooling strategies to test whether tropical adaptation changes linearly accessible climate signal before task-specific fine-tuning.

3.4 Incomplete-List Evaluation

We evaluate complete species lists and six simulated incompleteness settings: proportional retention of 75%, 50%, or 25% of the original species, and hard caps of 20, 10, or 5 species per cell. Retained species are sampled without replacement, and their relative order is preserved. Five truncation seeds, 101 through 105, define the same shortened inputs for all compared models. The hard-cap settings do not use abundance information because the stored assemblages contain species identities only. A0-FT is evaluated on complete lists only.

3.5 Retraining-Based Spatial Partition Sensitivity

To test sensitivity to the original partition, we construct a new spatial-block split without using the previous train/test assignment. We take the same 1,077 labeled cells, set aside their original train/test assignment, and regroup the underlying 5 km grid into disjoint blocks of 8×8 cells, corresponding to a nominal 40 km block width (Figure 1). Using a fixed random seed, whole blocks rather than individual cells are assigned to the test set until roughly 30% of the data falls there, which yields 748 training and 329 test cells across 104 and 53 blocks respectively, with no block or cell shared between the two sides; after removing cells with missing climate values, 743 training and 327 test cells remain.

We retrain LightGBM, A0-FT, and all five seeds of A2-MT from scratch on this new split, with vocabulary expansion and domain-adaptive pretraining once again restricted to the new training cells only; we do not retrain A2-ST or the frozen probes on this split. For the incomplete-list evaluation on this split we use five of the seven original settings (complete list, 25%, and the three hard caps), and we freeze the exact list of which cells, sparsity levels, and truncation seeds go together into a single manifest of 6,867 rows, reused unchanged for both LightGBM and A2-MT so that the two models are always tested on exactly the same inputs.

The split-generation procedure is deterministic and versioned: it groups the 5 km lattice into nominal 8×8-cell blocks, randomly permutes whole

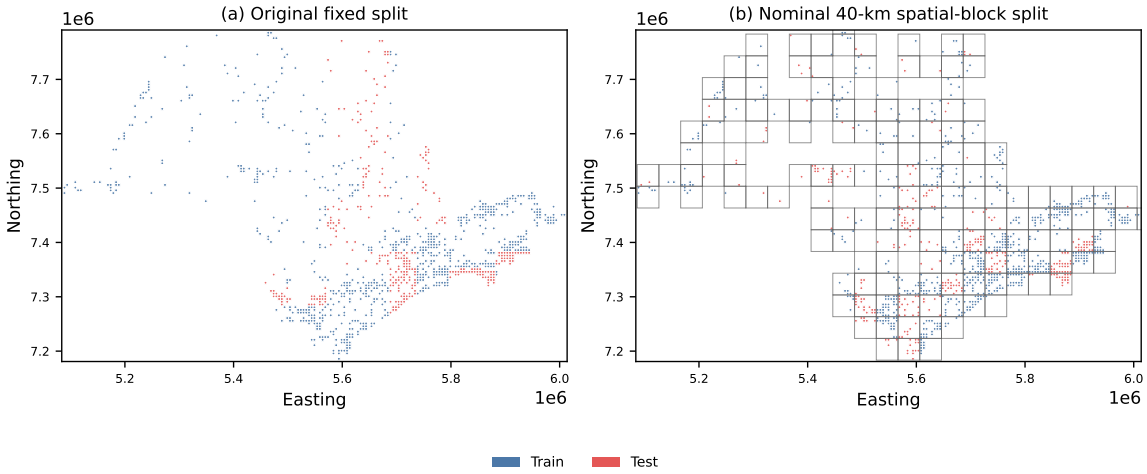


Figure 1: Spatial structure of the original fixed partition and the initial blocked retraining partition. (a) Original fixed train/test split. (b) Partition 1 of the nominal 40 km spatial-block design, shown as a representative example (Partitions 2-5 use the same blocking scheme; Section 5.3). Training and test cells are shown separately; panel (b) also shows the spatial blocks used to construct the partition. No cell or block is shared between training and test in panel (b).

blocks under the partition seed, and assigns blocks to the test set until the cumulative test-cell count first reaches or exceeds the 30% target. Split seeds, manifest hashes, and additional provenance are reported in Appendix D.

We refer to this split as Partition 1. To assess sensitivity to the specific block assignment, we generated four additional partitions (Partitions 2-5) under the identical nominal 40 km blocking design, using seeds drawn before inspecting outcomes on those four newly generated partitions (Appendix D). All five partitions reuse the same 1,077 cells; their test sets partially overlap, so we treat them as sensitivity analyses rather than independent replications (Section 5.3). Across Partitions 1-5, the raw train/test counts are 748/329, 736/341, 746/331, 752/325, and 743/334 cells, respectively, with test-block counts of 53, 55, 48, 37, and 40. All five partitions have zero train-test block overlap.

Exploratory vegetation-type classification and full training configurations are reported in Appendices E and D, respectively.

3.6 Evaluation and Uncertainty

Our main contrast is $\Delta R^2 = \overline{R^2_{\mathrm{A2-MT}}} - \overline{R^2_{\mathrm{LightGBM}}}$, where A2 values are averaged over training seeds and, for truncated conditions, over truncation realizations. The original-split analysis also reports a seed-ensemble estimand in the Appendix A. Uncertainty is assessed with 10,000-replicate hierarchical bootstrap procedures. For the retrained Partitions 1-5 analyses, each replicate resamples test cells, A2 training seeds, and truncation seeds with replacement; the sampled truncation realization is kept paired between A2-MT and LightGBM. We report 95% percentile intervals and two-sided add-one bootstrap tail areas. Benjamini-Hochberg adjustment is applied across the five completeness levels within each partition. Because the five partitions reuse cells and have overlapping test sets, we do not treat their 25 split-by-level contrasts as independent tests. Accordingly, cross-partition sign counts are treated only as descriptive summaries and are not assigned a combined significance test. Full definitions are provided in Appendix A.

4 Full-List Performance and Transfer

We first evaluate all models on complete species lists using the original fixed train/test split.

4.1 Full-List Regression

Species identity alone yields predictive performance for several climate targets. On complete test lists, the sparse LightGBM baseline reaches R^2=0.51 for BIO14 and R^2=0.48 for BIO12 (Table 1). However, most predictable variation is geographic: a LightGBM model using only centroid coordinates and elevation reaches R^2=0.88–0.98 across the four targets (Appendix G), and adding species features does not improve on geography alone. A partial Mantel analysis controlling for

geographic distance is nevertheless consistent with a small residual association between composition and climate for BIO1 and BIO14 (r=0.05–0.06, $p = 0.001$ for both BIO1 and BIO14, robust to two distance metrics), but not for BIO12. We therefore interpret species composition as carrying modest climate-related information beyond location, rather than as a substitute for geographic predictors; BIO14 is the target for which this residual signal also aligns with the subsequent incompleteness analysis.

The adapted A2-MT pipeline has higher five-seed mean R^2 than direct A0-FT by approximately 0.25 for BIO14 and 0.21 for BIO12. This A0-A2 contrast is descriptive; we did not run the hierarchical bootstrap or false-discovery-rate procedure for it. A2-MT also does not uniformly exceed LightGBM. BIO1 is highly seed-sensitive, BIO12 remains slightly below the baseline mean, and BIO15 is similar but variable. For BIO14, the A2-MT mean is 0.537 versus 0.510 for LightGBM, but the full-list hierarchical interval includes zero (CI_{95}=[−0.058, +0.117]). A2-ST reaches R^2=0.498 on the full BIO14 test set.

As an exploratory cross-task check, we also compare A0 and A2 on phytophysiognomy classification (Appendix E). A2 is descriptively higher than A0 on macro-F1, balanced accuracy, and accuracy, including a 0.10 absolute increase in accuracy. Relative to LightGBM, however, the comparison is mixed: A2 has higher accuracy but lower macro-F1 and balanced accuracy. Because this analysis uses only three seeds and no hierarchical bootstrap or multiplicity correction, we treat it as secondary evidence about A0-to-A2 transfer rather than evidence that the adapted pipeline is generally superior.

5 Performance under Incomplete Species Lists

We now assess performance under simulated species-list incompleteness, starting with the original split.

5.1 BIO14 under Severe Incompleteness (Original Split)

Figure 2 shows $\Delta R^2 = R^2_{\text{A2-MT}} - R^2_{\text{LightGBM}}$ for BIO14 across the seven list-retention settings on the original split. For the typical-pipeline summary, intervals include zero when lists retain about 30 or more species on average. Larger positive differences appear in the more severe settings that retain roughly 20 or fewer species (Table 5, Appendix B).

The result should not be interpreted as a universal 20-species threshold. The 25% and max-20 settings retain similar mean richness (15.5 and 17.3 species) through different mechanisms (proportional retention versus a hard cap), and their performance differs. The missingness mechanism therefore matters in addition to the number of retained species.

The typical-pipeline comparison survives within-family FDR correction at 25%, max-20, max-10, and max-5. The 25%, max-20, and max-5 settings also survive the global correction across all 84 tests. At the full, 75%, and 50% settings, the observed mean difference is positive but the hierarchical interval includes zero.

5.2 Spatial Sensitivity: Resampling vs. Retraining

We examined this pattern with respect to spatial variation using two complementary approaches, which yielded contrasting results.

The first check resamples the existing test set into spatial blocks at three scales (10, 20, 40 km, four offsets each; twelve partitions total), reusing the already-fitted models. The BIO14 advantage stayed positive across all twelve partitions at 25%, max-20, and max-5; max-10 was less stable (positive in eight of twelve). This reuses the original fitted models and test set, so it tests sensitivity to how that set is grouped, not whether the effect recurs under retraining on different data.

The retraining-based analysis addresses that question by fitting LightGBM, A0-FT, and all five A2-MT seeds on Partition 1, the initial nominal 40 km blocked assignment described in Section 3.5, and it tells a noticeably different story (Partition 1 row of Table 6, Appendix C). For BIO14, both models score lower on this split, and the full-list ranking seen originally reverses: where A2-MT's full-list mean had exceeded LightGBM's on the original split (0.537 against 0.510, Table 1, though that gap was not itself statistically supported), on the retrained split LightGBM reaches an R^2 of 0.399 for the same variable while A2-MT reaches only 0.329 on average across five seeds, and direct transfer without adaptation falls to -0.018. This reversal is informative on its own, since it shows that which model looks better on complete lists can flip between the original fixed split and Partition 1, independently of anything to do with sparsity.

On Partition 1, A2-MT remains descriptively

Table 1: Climate regression on full test lists ($n = 321$), original split. Language-model values are mean ± SD over five seeds; LightGBM is a single run. Bold marks the highest mean per target.

Model	BIO1	BIO12	BIO14	BIO15
LightGBM	**0.219**	**0.482**	0.510	**0.283**
A0-FT	-0.067 ± 0.077	0.255 ± 0.064	0.284 ± 0.047	0.101 ± 0.039
A2-MT	0.082 ± 0.169	0.461 ± 0.024	$\mathbf{0.537} \pm 0.028$	0.260 ± 0.085
A2-ST$_{\text{BIO14}}$	—	—	0.498 ± 0.062	—

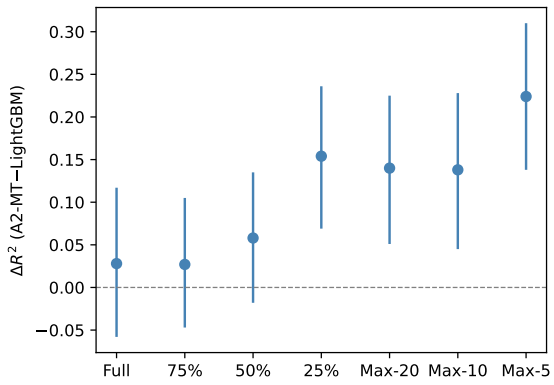


Figure 2: BIO14 performance difference (ΔR^2, A2-MT minus LightGBM) on the original fixed split, typical-pipeline summary. Points show observed ΔR^2; bars show 95% hierarchical bootstrap intervals. † marks significance after within-family and global correction; $*$ marks within-family correction only (Table 5, Appendix B).

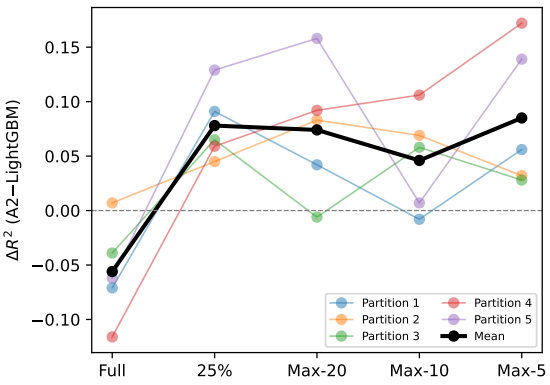


Figure 3: BIO14 ΔR^2 (A2-MT minus LightGBM) across the five 40-km spatial-block partitions. Thin lines are individual partitions; the thick black line is the cross-partition mean (Table 2). No level survives Benjamini-Hochberg correction in Partitions 1-3.

above LightGBM at 25% retention and max-20, but all five hierarchical confidence intervals include zero (Table 6), and no level survives Benjamini-Hochberg correction within the five-level family. The original BIO14 advantage did not retain inferential support on this alternative blocked partition: part of the descriptive pattern persists, but its statistical support does not.

5.3 Retraining Sensitivity Across Five Partitions

Figure 3 and Table 2 summarize BIO14 ΔR^2 across all five 40-km partitions (Partitions 1-5; per-partition detail in Appendix C).

Across complete lists, A2-MT exceeds LightGBM in only one of five partitions. Under incomplete lists, positive differences recur in all five partitions at 25% and Max-5, and in four of five at Max-20 and Max-10 (Table 2). These cross-partition summaries are descriptive because the partitions overlap.

Inferential support is less stable. No contrast survives within-partition BH adjustment in Partitions 1-3. In Partition 4, only Max-5 remains supported ($\Delta R^2 = +0.172$, 95% CI $[+0.061, +0.291]$, $q = 0.007$). In Partition 5, 25%, Max-20, and Max-5 remain supported (all $q = 0.016$). Thus, the direction of the incomplete-list effect is more stable than its inferential strength. Partitions 4 and 5, which retain adjusted contrasts, have the fewest test blocks (37 and 40, versus 48–55) and higher BIO14 test-set variance (SD 28.0–28.3 versus 21.5–23.1). For a fixed absolute prediction error, greater target variance can increase R^2 because the total sum of squares is larger. With only five overlapping partitions, however, test-block count and target variance are confounded, and we cannot attribute the difference to either factor. We did not identify a partition characteristic that cleanly explains the difference.

5.4 Other Climate Targets

No comparable typical-pipeline pattern is observed consistently for BIO1, BIO12, or BIO15. BIO12 shows isolated seed-ensemble advantages under some incomplete-list conditions on the original split, but this result depends on ensembling and is treated as secondary. BIO14 is therefore the

Table 2: Descriptive BIO14 ΔR^2 (A2-MT−LightGBM) across five 40-km partitions. Partitions overlap; not independent replications.

Level	Mean	Median	Range	A2 higher
Full	−0.056	−0.062	[−0.116, +0.007]	1/5
25%	+0.078	+0.065	[+0.045, +0.129]	5/5
Max-20	+0.074	+0.083	[−0.006, +0.158]	4/5
Max-10	+0.046	+0.058	[−0.008, +0.106]	4/5
Max-5	+0.085	+0.056	[+0.028, +0.172]	5/5

only target showing a repeated typical-pipeline difference across multiple severe-incompleteness settings. In particular, BIO1 performance is weak for both model families, so the present setup should not be interpreted as a successful temperature-reconstruction model; BIO15 likewise shows no consistent A2-MT advantage over LightGBM.

6 Additional Robustness Checks

The BIO14 analysis motivated six additional robustness and diagnostic checks targeting potential confounds and mechanisms not addressed above. Table 3 summarizes verdicts; full detail is in Appendices H–M.

Table 3: Summary of additional robustness and diagnostic checks related to the BIO14 results (original split unless noted).

Check	Outcome
Sequence length (64→128 tokens)	Same qualitative verdict at all 7 levels (Appendix H)
Vocabulary vs. DAPT ablation	Both components improve over A0 descriptively; combined gain is roughly additive (Appendix I)
Frequency-weighted removal	25% and max-20 lose significance; max-10 and max-5 survive (Appendix J)
Symmetric truncation exposure	LightGBM is higher at 5/7 levels; A2-MT is higher only at Max-10 and Max-5 (Appendix K)
Block-level bootstrap (Partitions 4-5)	All 4 previously supported contrasts retain positive intervals, with wider CIs (Appendix L)
Spatial buffer (10 km exclusion)	A2-MT's advantage over A0-FT widens, not narrows (Appendix M)

The most consequential of these is the symmetric-exposure check. LightGBM's original training used complete species-presence vectors, while A2's domain-adaptive pretraining used five random species-order reorderings per cell as augmentation. These reorderings change sequence order but do not expose A2 to truncated species lists. Thus, neither original pipeline was explicitly trained on the incomplete-list conditions used at evaluation. The symmetric-exposure experiment addresses this issue directly by retraining both model families with truncated training examples. Retraining LightGBM on a 23,219-row truncation-augmented training set (Appendix K) raises its BIO14 R^2 across all seven levels, exceeding A2-MT at full, 75%, 50%, 25%, and max-20; A2-MT remains descriptively higher only at Max-10 and Max-5. A size-matched control trained on bootstrap-resampled complete rows is lower than truncation-augmented LightGBM at every level for which the control was run, supporting a truncation-exposure effect rather than a sample-size explanation. This narrows the severe-sparsity claim to the two most extreme incompleteness regimes.

7 Discussion

Two questions naturally follow from a result that is specific to one variable and one range of list incompleteness: why that variable, and why that range. Precipitation of the driest month describes how severe the dry season gets, which plausibly connects more directly to seasonal water limitation (Tao et al., 2022) than annual mean temperature does, and that connection is easy to believe in a region that spans contrasting vegetation types. It is not, however, something our experiments establish; annual mean temperature also happens to vary less, relative to its own scale, in this dataset than precipitation of the driest month does, and altitudinal gradients or non-linear relationships between temperature and species composition could contribute as well. We treat the fact that this particular variable stood out as an empirical pattern worth reporting, not as a resolved ecological mechanism.

As noted in Section 1, this is a pattern we noticed rather than one we set out to confirm. The additional robustness checks also narrow how it should be interpreted. Frequency-weighted removal attenuates the advantage at 25% retention and Max-20, showing that the missingness mechanism matters, while symmetric truncation exposure substantially improves LightGBM and leaves A2-MT higher

only at Max-10 and Max-5. Thus, part of the original severe-sparsity pattern reflects how the models respond to a distribution shift they were not explicitly trained on, rather than architecture alone. A remaining possibility is that the adapted model uses broader co-occurrence structure when very few species remain, whereas LightGBM relies more heavily on individual species indicators, but we have not tested that mechanism directly. Consistent with this uncertainty, frozen A0/A2 representations yield $R^2 < 0$ under all three ridge-probe pooling strategies tested. We therefore treat the mechanism behind the residual Max-10 and Max-5 differences as unresolved.

A third, more consequential question is how much the choice of split itself matters. The original split's BIO14 advantage passed several within-split checks, including hierarchical bootstrap, multiplicity correction, and post-hoc spatial resampling of the same held-out set. None of these checks involves retraining on different train/test assignments. Partition 1 therefore provided an important warning: once both models were retrained under a nominal 40 km blocked assignment, the full-list ranking reversed and no incomplete-list contrast survived within-partition correction.

Post-selection regression to the mean is a complementary, non-exclusive explanation: because BIO14 was selected after inspection of the broader original-split results, some attenuation under new partitions is compatible with regression to the mean even without a systematic split effect. Partitions 2-5 do not resolve this ambiguity. The adjusted contrasts are concentrated in Partitions 4 and 5, which also have fewer test blocks and greater BIO14 test-set variance than Partitions 1-3; with only five overlapping partitions, these properties are confounded and do not provide a causal explanation for the difference. A block-level bootstrap on Partitions 4 and 5 leaves all four previously supported contrasts with positive intervals, although the intervals widen substantially (Appendix L), suggesting that the surviving results are not solely an artifact of cell-level resampling. Separately, a 10 km exclusion buffer applied to Partition 1 widens the BIO14 difference between A2-MT and A0-FT because direct transfer deteriorates more strongly (Appendix M). This buffered check concerns adaptation versus direct transfer rather than the A2-MT-versus-LightGBM sparsity comparison, so we treat it as complementary rather than decisive evidence about the headline result.

Putting the original split and the five retrained partitions side by side leads to a fairly modest practical conclusion. On the original split, the adapted pipeline showed a positive BIO14 difference in several severely incomplete settings, with longer observed training times in our heterogeneous runs, although these timings are not hardware-normalized and should not be interpreted as a controlled efficiency comparison (Table 4).

The retrained partitions narrow that considerably: descriptive advantages recur under severe incompleteness, but adjusted contrasts survive Benjamini-Hochberg correction in only two of five partitions (Partition 4 at Max-5; Partition 5 at 25%, Max-20, and Max-5), while Partition 1 shows no surviving contrast at any level. Given this uncertainty, we currently see the simpler LightGBM baseline as the more defensible default; the original-split result alone does not justify the additional complexity of the adapted pipeline. We also do not yet know how often real biodiversity-monitoring programs encounter list incompleteness as severe as the settings simulated here, or whether the pattern would persist across additional blocked partitions or other regions.

8 Conclusions

The combined A2 adaptation pipeline is descriptively higher than A0 across the evaluated regression and exploratory classification metrics, but it does not uniformly outperform LightGBM. Across five alternative 40 km spatial-block partitions, A2-MT shows no consistent full-list advantage, and the BIO14 incomplete-list effect remains partition-dependent.

The strongest practical claim narrows further under symmetric truncation exposure: on the original split, A2-MT remains higher than LightGBM only at Max-10 and Max-5 after both model families are trained on truncated examples. We therefore view the adapted pipeline as potentially useful under extreme list incompleteness rather than as a general replacement for the simpler baseline. More broadly, the results show that within-split robustness checks do not substitute for retraining across alternative spatial partitions.

Limitations

All experiments concern São Paulo state, so transfer to other tropical regions remains untested. The original fixed split was inherited from an earlier

Table 4: Descriptive wall-clock comparison, original split. LightGBM: four independent regressors, serial sum. A2-MT DAPT+FT includes DAPT; FT-only excludes it. Values with ± report mean ± SD over five seeds. Hardware was heterogeneous and was not systematically recorded for LightGBM; these wall-clock times are therefore descriptive and not directly comparable across model families, so no training-time ratio is reported.

Model	Train (s)	Infer (s)
LightGBM (4 targets)	21.4	0.048
A0-FT (GPU)	62.8 ± 6.3	0.51 ± 0.10
A2-MT FT only (GPU)	66.7 ± 3.9	0.48 ± 0.03
A2-MT DAPT+FT	210.2 ± 9.0	0.48 ± 0.03

phase of the project, and its exact generation procedure could not be reconstructed from the retained artifacts. Partitions 1-5 were therefore generated under an explicit, versioned spatial-block procedure, but they reuse the same 1,077 cells and have partially overlapping test sets; they should be interpreted as sensitivity analyses rather than independent replications. BIO14 was also selected post hoc after inspection of the broader original-split results, so selection-induced effect inflation remains possible.

The nominal 40 km partitions have zero train–test block overlap but impose no minimum physical separation across block boundaries. A 10 km buffered sensitivity analysis suggests that residual proximity does not explain the A2-versus-A0 transfer difference on the tested partition, but this check used only three seeds, one buffer distance, and did not directly test the A2-MT-versus-LightGBM sparsity comparison.

The primary Partitions 1-5 bootstrap resamples individual test cells rather than spatial blocks. A targeted block-level bootstrap on Partitions 4 and 5 leaves all four previously supported contrasts with positive intervals, but with wider uncertainty; we did not repeat block-level inference for every partition and condition.

Species sequences are stored in lexicographic rather than abundance-oriented order, creating a mismatch with Pl@ntBERT pretraining. The original incompleteness analysis uses uniform random removal, whereas real biodiversity records have structured detection and sampling biases. A frequency-weighted sensitivity analysis attenuates several of the original differences, showing that the results are sensitive to the missingness mechanism. Likewise, symmetric truncation exposure substantially improves LightGBM and narrows the residual A2-MT advantage to Max-10 and Max-5 on the original split. Finally, LightGBM uses a train-fitted species feature space, taxa appearing only at test time are not represented as learned baseline features, and recorded wall-clock times were obtained on heterogeneous hardware and are descriptive rather than hardware-normalized efficiency estimates.

Acknowledgements

This work was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21].

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. https://aclanthology.org/N19-1423.

Stephen E. Fick and Robert J. Hijmans. 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315. https://doi.org/10.1002/joc.5086.

GBIF.org. 2026. GBIF occurrence download. https://doi.org/10.15468/dl.evgrnx. Accessed 7 May 2026.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. https://aclanthology.org/2020.acl-main.740.

Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. 2021. Sampling biases shape our view of the natural world. *Ecography*, 44(9):1259–1269. https://doi.org/10.1111/ecog.05926.

John Imbrie and Nilva G. Kipp. 1971. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. In Karl K. Turekian, editor, *The Late Cenozoic Glacial Ages*, pages 71–181. Yale University Press, New Haven.

Instituto Florestal. 2020. Inventário Florestal do Estado de São Paulo. Technical report, Secretaria de Infraestrutura e Meio Ambiente do Estado de São Paulo, São Paulo, SP, Brazil.

César Leblanc, Pierre Bonnet, Maximilien Servajean, and Alexis Joly. 2024. Pl@ntBERT: Leveraging Large Language Models to Enhance Vegetation Classification through Species Composition Analysis. In *European Congress of Conservation Biology (ECCB) - Book of Abstracts*, page 121.

César Leblanc, Pierre Bonnet, Maximilien Servajean, Wilfried Thuiller, Milan Chytrỳ, Svetlana Aćić, Olivier Argagnon, Idoia Biurrun, Gianmaria Bonari, Helge Bruelheide, and 1 others. 2025. Learning the syntax of plant assemblages. *Nature Plants*, 11(10):2026–2040.

Volker Mosbrugger and Torsten Utescher. 1997. The coexistence approach—a method for quantitative reconstructions of Tertiary terrestrial palaeoclimate data using plant fossils. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 134:61–86. https://doi.org/10.1016/S0031-0182(96)00154-X.

Jonathan T. Overpeck, Thompson Webb, and I. Colin Prentice. 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research*, 23(1):87–108.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. 2020. Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nature Communications*, 11:4540. https://doi.org/10.1038/s41467-020-18321-y.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography*, 40(8):913–929. https://doi.org/10.1111/ecog.02881.

Shengli Tao, Jérôme Chave, Pierre-Louis Frison, Thuy Le Toan, Philippe Ciais, Jingyun Fang, Jean-Pierre Wigneron, Maurizio Santoro, Hui Yang, Xiaojun Li, Nicolas Labrière, and Sassan Saatchi. 2022. Increasing and Widespread Vulnerability of Intact Tropical Rainforests to Repeated Droughts. *Proceedings of the National Academy of Sciences*, 119(37):e2116626119. https://doi.org/10.1073/pnas.2116626119.

A Statistical Methodology Detail

Estimands. *Typical-pipeline*: within each bootstrap replicate, average the seed-specific ΔR^2 over the resampled five-seed multiset (captures mean pipeline performance under optimization-seed uncertainty). *Seed-ensemble*: within each replicate, average predictions over the same resampled multiset before computing a single ΔR^2 (captures an ensembled-deployment estimand). For truncated levels, reported "observed" values are typical-pipeline, further averaged across the five truncation-seed scenarios.

Hierarchical bootstrap. For the original fixed-split analysis, 10,000 replicates resample test cells and model seeds with replacement; truncation scenarios are treated as matched repeated conditions. For the retrained Partitions 1-5 analysis, each replicate resamples test cells, A2 model seeds, and truncation seeds with replacement, while preserving the sampled truncation realization between A2-MT and LightGBM. CIs are 95% percentile intervals.

Multiplicity. Benjamini-Hochberg FDR (Benjamini and Hochberg, 1995) is applied separately within six specified analysis-by-estimand families: MT vs. LightGBM under the typical-pipeline and seed-ensemble estimands (28 tests each, spanning four targets and seven sparsity levels), and ST vs. LightGBM and ST vs. MT under each estimand (7 BIO14 tests per family) - 84 tests total, all on the original split. We additionally apply a global BH correction across all 84 tests as a stricter sensitivity analysis.

Bootstrap tail areas. Two-sided finite-bootstrap tail areas use the add-one correction, with n_+/n_- the counts of replicates where ΔR^2 falls on each

side of zero and B the total number of replicates:

$$p_+ = \frac{n_+ + 1}{B + 1}, \quad p_- = \frac{n_- + 1}{B + 1}, \tag{1}$$

$$p_{\text{two}} = \min(1,\ 2\min(p_+, p_-)). \tag{2}$$

We use these finite-bootstrap tail areas as inputs to the BH procedure; because they are bootstrap-derived rather than parametric p-values, we interpret the adjusted values jointly with the percentile intervals and effect sizes rather than as standalone evidence. On the original split, this two-sided construction is used throughout. For Partitions 1-5, the same two-sided add-one construction is applied directly to the retained bootstrap replicate arrays for all five completeness levels in each partition, followed by Benjamini-Hochberg adjustment within each partition (Table 6).

B Full Hierarchical Bootstrap Results (Original Split)

Table 5: Hierarchical bootstrap: BIO14, A2-MT typical pipeline vs. LightGBM, original split (321 test cells). ΔR^2=A2-MT−LightGBM. CI: 95% percentile interval, 10,000 replicates. q_{BH}: within-family; † also survives global-84; $*$ within-family FDR only (does not survive global-84). ">0.05": within-family FDR family but not a discovery at 0.05. "Clear": 95% CI entirely above zero. "Inconclusive": CI includes zero.

Level	Spp.	ΔR^2	95% CI	q	Verdict
Full	60.7	+0.028	[−0.058, +0.117]	>0.05	inconclusive
75%	45.8	+0.027	[−0.047, +0.105]	>0.05	inconclusive
50%	30.6	+0.058	[−0.018, +0.135]	>0.05	inconclusive
25%	15.5	+0.154	[+0.069, +0.236]	0.013	**clear**†
Max-20	17.3	+0.140	[+0.051, +0.225]	0.013	**clear**†
Max-10	9.8	+0.138	[+0.045, +0.228]	0.042	**clear**$*$
Max-5	5.0	+0.224	[+0.138, +0.310]	0.006	**clear**†

For BIO1 and BIO15 we also counted, per individual seed and sparsity level, how often the seed-specific ΔR^2 bootstrap interval excluded zero ("decisive"; all other cases were treated as inconclusive). Of the decisive seed×level comparisons (14 for BIO1 and 8 for BIO15), BIO1 favors LightGBM in 13 and A2 in 1; BIO15 favors LightGBM in 6 and A2 in 2.

C Multisplit BIO14 Results

"Clear" denotes a positive 95% bootstrap interval together with $q < 0.05$ after Benjamini-Hochberg adjustment within the corresponding partition; "inconclusive" denotes all remaining cases.

Table 6: BIO14 ΔR^2=A2-MT−LightGBM by partition and list-completeness level. CIs are 95% hierarchical bootstrap percentile intervals; q is Benjamini-Hochberg adjusted within each partition across the five evaluated levels.

Partition	Level	ΔR^2	95% CI	q	Verdict
1	Full	−0.071	[−0.203, +0.052]	0.642	inconclusive
1	25%	+0.091	[−0.027, +0.201]	0.642	inconclusive
1	Max-20	+0.042	[−0.083, +0.159]	0.642	inconclusive
1	Max-10	−0.008	[−0.145, +0.110]	0.894	inconclusive
1	Max-5	+0.056	[−0.100, +0.191]	0.642	inconclusive
2	Full	+0.007	[−0.129, +0.136]	0.904	inconclusive
2	25%	+0.045	[−0.059, +0.139]	0.615	inconclusive
2	Max-20	+0.083	[−0.032, +0.184]	0.584	inconclusive
2	Max-10	+0.069	[−0.052, +0.172]	0.584	inconclusive
2	Max-5	+0.032	[−0.064, +0.117]	0.615	inconclusive
3	Full	−0.039	[−0.246, +0.138]	0.923	inconclusive
3	25%	+0.065	[−0.100, +0.194]	0.923	inconclusive
3	Max-20	−0.006	[−0.196, +0.136]	0.961	inconclusive
3	Max-10	+0.058	[−0.122, +0.194]	0.923	inconclusive
3	Max-5	+0.028	[−0.144, +0.162]	0.923	inconclusive
4	Full	−0.116	[−0.303, +0.063]	0.430	inconclusive
4	25%	+0.059	[−0.101, +0.211]	0.470	inconclusive
4	Max-20	+0.092	[−0.096, +0.272]	0.430	inconclusive
4	Max-10	+0.106	[−0.077, +0.297]	0.430	inconclusive
4	Max-5	**+0.172**	**[+0.061, +0.291]**	**0.007**	**clear**
5	Full	−0.062	[−0.184, +0.068]	0.409	inconclusive
5	25%	**+0.129**	**[+0.032, +0.226]**	**0.016**	**clear**
5	Max-20	**+0.158**	**[+0.043, +0.285]**	**0.016**	**clear**
5	Max-10	+0.007	[−0.129, +0.157]	0.931	inconclusive
5	Max-5	**+0.139**	**[+0.032, +0.247]**	**0.016**	**clear**

D Verified Training Configuration

Table 7 summarizes settings recovered from retained training scripts, per-seed metrics, resolved trainer configurations, and serialized LightGBM models. For DAPT, seeds 42-44 retain core settings in per-seed metrics and the training script; seeds 45-46 additionally preserve resolved trainer configurations linked to the same script SHA-256.

Partition provenance. Partitions 2-5 use four seeds before inspecting model outcomes on those four newly generated partitions; Partition 1 (seed 20260706) had already been used as the initial retraining-based spatial sensitivity check. The base split generator and its wrapper script are content-hashed. Frozen sparsity-manifest SHA-256 values for Partitions 1-5 are `4a2efdbc...476abee7`, `53e5ca64...66c3b`, `35955806...36dc`, `b5d560ee...c5f6`, and `6517ede6...d6b2`, respectively; full manifest and generator hashes are retained in the accompanying repository.

Model and tokenizer. All transformer runs start from the public Pl@ntBERT masked-language checkpoint (`CesarLeblanc/plantbert_fill_mask_model`), using its original tokenizer and base vocabulary of

44,708 tokens. Tropical adaptation adds whole-species tokens observed in the training data; species with at least two training-corpus occurrences are eligible. Of 5,253 requested tokens, 5,108 were added (145 were already present in the base vocabulary), yielding a final vocabulary of 49,816 tokens. New-token embeddings are randomly initialized and subsequently updated during DAPT and downstream fine-tuning; in the vocabulary-only ablation they receive no DAPT adaptation and are learned only during downstream fine-tuning. Hyperparameters in Table 7 were fixed in advance and were not selected using the reported test set; no early stopping or test-driven hyperparameter selection was performed.

Table 7: Verified configuration used in the reported experiments.

Component	Setting
FT backbone	Fully trainable
FT pooling	CLS representation
FT head	Linear regression
FT outputs	4 (MT), 1 (ST)
FT target scaling	StandardScaler, train-only fit
FT loss	MSE
FT epochs / batch	10 / 8
FT learning rate	10^{-5}
FT optimizer	AdamW
FT max sequence length	64
FT precision / seeds	FP16 / 42-46
DAPT objective	Masked language modeling
DAPT epochs / batch	3 / 8
DAPT learning rate	5×10^{-5}
DAPT max sequence length	64
DAPT masking probability	0.15
DAPT augmentation	5 random species-order reorderings per cell
DAPT optimizer	AdamW; fused in resolved seed-45/46 configs
DAPT scheduler / weight decay	Linear / 0, in resolved seed-45/46 configs
LightGBM boosting	GBDT
LightGBM leaves / max depth	31 / -1
LightGBM learning rate	0.05
LightGBM estimators	300
LightGBM row / feature subsampling	1.0 / 1.0
LightGBM L1 / L2	0 / 0
LightGBM random state	42

E Phytophysiognomy Classification (Exploratory)

No hierarchical bootstrap or multiplicity correction was applied to the classification comparisons as follows; treat as exploratory secondary evidence only. Evaluated on the original fixed split only.

Table 8: Phytophysiognomy classification, fixed test set (326 cells; unlike regression, classification does not require complete climate targets, hence 326 rather than 321). A0/A2: mean ± SD over 3 seeds. Bold: highest mean per column.

Model	Macro-F1	Bal. Acc.	Acc.
LightGBM	**0.449**	**0.441**	0.727
A0 direct	0.337 ± 0.021	0.340 ± 0.026	0.645 ± 0.038
A2 (vocab.+DAPT)	0.374 ± 0.020	0.380 ± 0.030	$\mathbf{0.745} \pm 0.035$

Per-class F1 (LightGBM): FOD 0.834, FES 0.622, FOM 0.214, Formação Pioneira 0.125, showing markedly weaker performance on the two rarer classes.

F Geographic Border-Cell Check

Of 1,077 IFSP30 cells, 113 (10.5%) have centroids outside São Paulo's political boundary (maximum 3.5 km, median 950 m), distributed as 72 training and 41 test cells. Excluding the 36 affected test cells present in the 321-cell climate evaluation subset and re-running the hierarchical bootstrap for BIO14 leaves 25%, max-20, and max-5 clear (q=0.025, 0.037, 0.006); max-10 becomes inconclusive (ΔR^2=0.129, q=0.057). A placebo control of 20 independent random 36-cell exclusions shows this is a general small-sample effect, not specific to the border cells: 11 of 20 random exclusions similarly push max-10 above the 0.05 threshold, and the border-cell exclusion's ΔR^2 falls at the median of the placebo distribution.

G Geographic Control

A LightGBM model using only cell centroid coordinates and elevation reaches R^2=0.876 (BIO1), 0.903 (BIO12), 0.904 (BIO14), and 0.976 (BIO15) on the 321-cell test set. Combining species features with these three geographic features does not improve over geography alone; geographic features account for 98–99% of total split gain in the combined model. Elevation was obtained via the OpenTopoData SRTM90m API for all 1,070 cells with complete climate targets. A partial Mantel test (composition vs. climate, controlling for geography; 571,915 pairs, 999 permutations) gives r=0.060 (p=0.001) for BIO1 and r=0.048 (p=0.001) for BIO14, but r=0.006 (p=0.533) for BIO12; both results are essentially unchanged

(third decimal place) under Sørensen-Dice in place of Jaccard distance.

H Sequence-Length Confound Check

Fine-tuning uses a 64-token maximum context; given mean richness of 58.9 species, this truncates 57.6%, 48.0%, 34.1%, and 14.1% of test cells at full, 75%, 50%, and 25% retention, falling below 9% at all three hard-cap levels. We retrained the full A2-MT pipeline (five seeds) with 128-token context, reducing truncation substantially (e.g., 57.6%→26.8% at full). This run also used an independently reconstructed, larger DAPT corpus, so it is a robustness check on the truncation confound rather than an isolated sequence-length ablation; we did not have a 128-token run on the exact original DAPT corpus. The qualitative pattern is unchanged at all seven levels: no significant advantage at full/75%/50%, and a within-family-FDR-surviving advantage at 25%, max-20, max-10, and max-5, matching Table 5 exactly in verdict.

I Vocabulary Expansion vs. DAPT Ablation

We trained two variants isolating each component of tropical adaptation, three seeds each, on the original split. *DAPT-only* (unexpanded vocabulary, DAPT on training cells) reaches BIO14 R^2=0.375±0.036. *Vocabulary-only* (expanded vocabulary, no DAPT; new token embeddings are randomly initialized and updated only during downstream fine-tuning) reaches R^2=0.439±0.131. Both improve over A0-FT, recomputed here over the same three seeds as the ablation arms (R^2=0.263±0.034; Table 1 reports 0.284 over all five seeds); combined, A2-MT reaches 0.537±0.028. The sum of individual improvements (+0.112 and +0.176, totaling +0.288) closely matches the combined improvement (+0.274), suggesting the two components contribute largely independently. Vocabulary-only's standard deviation is roughly four times larger than the other variants', consistent with ∼5,108 randomly initialized new-token embeddings having to be learned from only 749 downstream training cells without prior DAPT adaptation; with three seeds we do not treat the vocabulary-vs-DAPT ranking as statistically established.

J Frequency-Weighted Removal

Uniform random truncation does not reflect structured sampling and recording biases in GBIF (Hughes et al., 2021). We therefore built an alternative manifest retaining species with probability weighted by their training-corpus record frequency (common species favored to survive truncation) and re-ran inference with the existing LightGBM and five-seed A2-MT models. Training-corpus record frequency is used here as a simple proxy for structured recording probability and does not separate ecological commonness from true detectability. Under a self-contained six-level FDR family, 25% and max-20 no longer reach significance (q=0.78, 0.09); max-10 and max-5 remain significant (q=0.028, 0.001). ΔR^2 decreases at every level (e.g., 25%: 0.154→0.025; max-5: 0.224→0.172).

K Symmetric Truncation Exposure

LightGBM's original training never saw truncated inputs. We retrained LightGBM on a 23,219-row truncation-augmented training set (749 original rows plus 6 levels × 5 truncation seeds of resampled training cells) and separately retrained A2-MT (five seeds) from its original DAPT checkpoints on the same augmented data. A size-matched control (LightGBM trained on 23,219 bootstrap-resampled *untruncated* rows) isolates whether gains reflect truncation exposure or sample size alone.

Table 9: BIO14 R^2 under symmetric truncation exposure, original split.

Level	LGBM aug.	LGBM ctrl.	A2-MT aug.
Full	0.665	0.577	0.511±0.048
75%	0.636	–	0.503±0.042
50%	0.567	–	0.443±0.035
25%	0.435	0.169	0.361±0.069
Max-20	0.525	0.269	0.458±0.057
Max-10	0.338	0.048	0.373±0.046
Max-5	0.161	−0.148	0.177±0.068

The control supports an exposure effect rather than a sample-size explanation: at Max-5, the control (untruncated, resampled) falls *below* zero, while augmented LightGBM stays positive, showing training on truncated examples specifically helps at severe sparsity. LightGBM-augmented exceeds A2-MT-augmented at five of seven levels; A2-MT-augmented is higher only at max-10 (+0.035) and max-5 (+0.016).

L Block-Level Bootstrap (Partitions 4-5)

The main multisplit bootstrap resamples individual cells; Partitions 4-5, where all surviving contrasts concentrate, have the fewest test blocks (37, 40) among the five partitions. Re-running the hierarchical bootstrap by resampling whole spatial blocks leaves all four previously supported contrasts with positive intervals, though with wider intervals (e.g., Partition 4 Max-5: CI_{95}=[0.034, 0.320], versus a much tighter cell-level interval).

M Spatial Exclusion Buffer

We constructed a partition using Partition 1's block assignment but additionally excluding training cells within 10 km (Euclidean) of any test cell, reducing training from 748 to 628 cells. Retrained on this buffered split (three seeds), A0-FT's BIO14 R^2 falls to -0.153 ± 0.103 (from ≈ 0.27 on the original split), while A2-MT falls to 0.310 ± 0.129 (from 0.537). The relative advantage of adaptation over direct transfer widens rather than narrows under the buffer, suggesting A0-FT's original-split performance relied more on spatial proximity than A2-MT's did. Three seeds and a single buffer distance limit how much weight this result can bear.

Extracting information on public water shortages using news media and LLMs: a pilot study in Sweden

Jeanne Fernandez[1,2] **Shorouq Zahra**[1,3] **Mariana Madruga de Brito**[4] **Johanna Mård**[1,2]

[1]Uppsala University [2]Centre of Natural Hazards and Disaster Science
[3]RISE Research Institutes of Sweden [4]Helmholtz Centre for Environmental Research - UFZ

Abstract

Slow-onset events such as droughts lack a clearly defined onset, and systematic databases do not capture their impacts at fine scales. News reports can help fill this gap by providing information on how drought-driven water shortages affect communities and how authorities respond. Yet, extracting structured information from text remains challenging. Here, we use news articles and official public communications on water-saving recommendations and water-use restrictions in Sweden as a case study to benchmark LLMs for drought-related information extraction. We compare the proprietary GPT-5 series against three large open-access models (Llama3.3, Deepseek, and GPT-OSS) in extracting variables such as the date, location, and causes of water-saving measures. Using prompt engineering, structured outputs, and minimal post-processing, we evaluate performance using a test set of 210 manually annotated articles. The GPT-5 model series outperforms all open-access models across all variables. We find that, without task-specific fine-tuning, open-source models are sometimes outperformed by a keyword-based baseline, suggesting that general-purpose capabilities alone are insufficient for reliably extracting structured data. These findings underscore the challenges of monitoring water-shortage risks based on texts, particularly when extracting complex variables.

1 Introduction

Drought is a slow-onset event expected to become more frequent in many regions due to climate change (Chiang et al., 2021), requiring novel approaches to monitor its progress (Wens et al., 2026). While meteorological and hydrological indicators can characterize the physical hazard, information on impacts and response measures is fragmented across unstructured text sources. This challenge is pronounced in regions considered water-rich, such as Scandinavia, where the emerging risk of water shortage is under-studied (Lindqvist et al., 2021). However, the 2018 European-scale drought marked a turning point in awareness of water scarcity, receiving widespread media coverage. In Sweden, news outlets and municipal websites disseminated information on water shortages, including water-saving recommendations and restrictions on non-essential water use. Although these measures are among the most common responses to water scarcity in Sweden (Teutschbein et al., 2023), they are not systematically recorded in regional or national databases. Thus, leveraging news aggregator datasets and text-mining approaches to record such measures over space and time can provide insights on societal responses to water scarcity and help document the reliability and resilience of drinking water systems in a changing climate.

However, extracting useful information from news requires more than identifying publication dates and municipality names. One must separate saving recommendations from mandatory restrictions, identify whether measures affect an entire municipality or only parts of it, and extract precise start and end dates for these measures. Also, to support downstream analyses, information extraction methods must identify the cause of such shortages, distinguishing physical drivers (e.g., precipitation deficits and increased evapotranspiration associated with high temperatures) from demand-driven factors (e.g., high water consumption).

A key challenge then becomes the journalistic style of news texts, making it difficult for conventional natural language processing (NLP) approaches to reliably assess the relationships among a location, a date, a measure, and causes. This is particularly problematic for texts with information about multiple co-occurring or sequential measures and covering multiple locations. In addition, the high semantic similarity in the formulation of mandatory water-use restrictions and water-saving recommendations makes it difficult to characterize

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 99–115
October 28, 2026 ©2026 Association for Computational Linguistics

the type of measure in place. Both measures may contain terms associated with water conservation. Finally, characterizing article relevance for the proposed analysis requires context understanding and cannot be achieved with simple classification methods.

Advances in large language models (LLMs) have opened new opportunities to extract structured information about climate events from texts. Their in-context understanding abilities are valuable for such tasks (Dong et al., 2024). Several studies have applied LLMs to build climate hazard databases from text data. For instance, Li et al. (2024) and Vela-Tambo et al. (2025) find that LLMs are promising tools for climate-related data extraction from reports and news articles, even when performing cross-lingual information extraction in the case of the latter.

Nonetheless, LLM performance on domain-specific tasks and on extracting subtle signals of drought risk, such as water-shortage measures, has not been evaluated. Moreover, open-access models are cheaper and facilitate replicability as they can be run locally, often under non-restrictive licenses. Despite being trained for a wide range of general tasks, their baseline performance on structured data extraction relative to proprietary models and keyword-based NLP methods is unclear.

We propose an LLM-based method to extract data on public water supply shortage risks in Sweden from news coverage of water savings and water-use restrictions (Figure 1). The overarching goal is to identify risk hotspots for water supply, i.e., places facing frequent and long-lasting restrictions, as well as the reasons why restrictions are applied. Specifically, the resulting spatiotemporal database can help identify patterns in municipal water-management responses to droughts and understand how these escalate from warnings to water-use restrictions (Figure 2). In combination with additional datasets (e.g., drought indicators, municipal water source characteristics), the database can contribute to a more holistic understanding of water shortage risks by linking drought hazards, water-system vulnerabilities, and implemented responses. This, in turn, can support drought preparedness and provide relevant insights for water managers and planners.

We benchmark the GPT-5 series proprietary models against three open-source models (Llama 3.3, Deepseek, GTP-OSS) and one hand-crafted rule-based baseline in extracting structured information from unstructured texts. The experimental setup relies on prompt engineering and limited post-processing to evaluate model performance under minimally tuned conditions. All prompts, code, and metadata are made available on GitHub.[1]

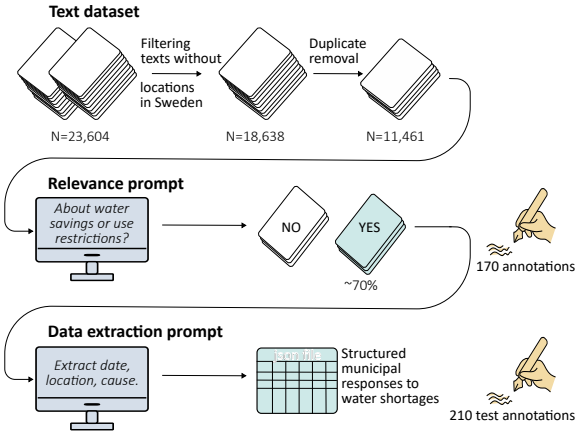


Figure 1: Data extraction pipeline and its evaluation

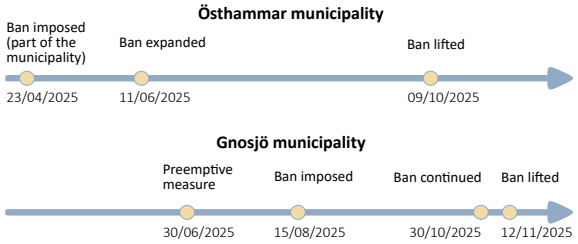


Figure 2: Example reconstruction of a timeline of measures for Östhammar and Gnosjö municipalities in 2025, based on information extracted from news articles with GPT-5.5.

2 Related work

News-based analysis of drought and water scarcity is a growing research field, showing that news can offer extensive, yet still underutilized, information on this hydrological extreme. Existing NLP applications have focused on detecting and classifying drought impacts into broad classes (Sodoge et al., 2023; Vela-Tambo et al., 2025). Studies have also linked trends in news coverage of droughts to domestic water use patterns, focusing on communication volume and sentiment analysis (Lee et al., 2026). More recently, embedding-based semantic similarity methods have been used to identify fuzzy concepts, such as drought-related groundwater conflicts, in news articles (Sodoge et al., 2026).

[1] https://github.com/jeannefdz/swedish_water_restrictions

More broadly, LLMs are increasingly used in climate research to extract structured information from unstructured text due to their contextual understanding and few-shot learning capabilities (Brown et al., 2020; Vaswani et al., 2017). Applications include the construction of climate impact databases (Li et al., 2024; Vela-Tambo et al., 2025; de Brito et al., 2026), monitoring societal impacts of droughts via social media (Wang et al., 2025), hazard fact-sheets and knowledge graphs (Ronco et al., 2026), or evidence of gaps in the regional coverage of hazards in scientific literature (Carvalho et al., 2025). We extend these applications by extracting reports on municipal drinking water saving recommendations and use restrictions from Swedish media texts, providing a structured dataset of drought response measures.

Despite the success of LLMs on some of these tasks, several factors influence their performance. Larger commercial models often outperform open-access models, although open-access models can match the performance of proprietary LLMs on some tasks (Li et al., 2024). Performance also depends on prompt design, with prompt length and detail needing to be adapted to model size (Vela-Tambo et al., 2025). Additionally, both Vela-Tambo et al. (2025) and Li et al. (2024) find that LLMs struggle to produce structurally valid outputs, with performance depending on model size and the effectiveness of supporting pre- and post-processing techniques such as schema validation.

3 Text corpus

The data used is based on a corpus of Swedish texts on municipal water-use restrictions (also referred to as "watering bans") and water-saving recommendations. The choice of data source is justified by ease of access, stability over time, and spatiotemporal coverage. While municipal websites represent an authoritative source regarding current restrictions, webscraping requires specific extraction scripts for each webpage structure and these structures evolve over time. In addition, older website versions containing historical restrictions may be archived inconsistently, making it challenging to retrieve data on past measures.

We retrieved articles[2] from Swedish sources in the Retriever media archive[3], a database with 4,574 Swedish language media sources[4], including printed press, web, and TV/Radio. We collected both editorial and non-editorial news articles and official public communications (e.g., press releases and public advisories) published between January 1st, 2000 and December 31st, 2025, using keyword searches of titles and headlines. The search terms, informed by similar searches (Quesnel and Ajami, 2017) and expert knowledge, comprise four separate queries (Appendix A, Table 6). These queries were developed by testing alternative combinations of synonyms and expressions for watering bans and water savings and checking the effect of additional keywords on the number of articles returned (Appendix A). This strategy maximized recall, resulting in a corpus of 23,604 texts.

The subsequent pre-processing steps involved cleaning the texts for more efficient NLP processing, removing articles that mentioned no location in Sweden, and clustering duplicates (Figure 1). To reflect municipal-level management of public water supply in Sweden, we only included articles that mentioned either a municipality (i.e., local administrative unit level 2 (LAU 2)) or a statistically defined urban area (*"Tätort"* in Swedish). An urban area, as defined by Statistics Sweden, is a continuous built-up area with at least 200 inhabitants that meets specific distance criteria between buildings (Statistikmyndigheten SCB, 2023). We excluded articles that only referred to larger or smaller administrative units (e.g., regions, counties, districts, neighborhoods) as these introduced noise in the dataset. We used regular expressions (regex) patterns to search the articles' main text for municipality and urban area names, including their lexical variations (see Appendix A for details). This simple filtering was found to perform better than more advanced named entity recognition methods for Swedish locations, based on manual checking of some articles. In this step, 4,966 articles were removed.

Duplicate and near-duplicate articles were identified using Term-Frequency Inverse Document Frequency (TF-IDF) and cosine similarities. This approach weighs the relative importance of each term in the corpus, giving a higher weight to terms that are frequent in a document but rare in the overall corpus. It ensures that the most informative

[2] In this paper, we use "article" as a generic term for records retrieved from the Retriever media archive. Most records originate from printed and online news media but the corpus also includes municipal communications and press-releases.

[3] https://retriever.se

[4] Retriever search interface, March 2026

words are given more weight and can capture partial matches between paraphrased journalistic texts (Rajaraman and Ullman, 2011). To build the TF-IDF matrix, we used 5- to 8-character n-grams. The cut-off value for defining duplicates was set to 0.8 (1 representing full cosine similarity) as many texts in our corpus are near duplicates with identical text bodies but paraphrased titles, leads, captions, or closing paragraphs. The choice of a high cut-off value is consistent with other duplicate-detection studies (de Brito et al., 2025; Madureira et al., 2026) and was validated by examining the distributions of cosine similarities (Appendix A, Figure 4) and duplicate clusters for a subset of 300 articles. The final corpus contains 11,461 unique texts. A quick sampling of these texts showed that most were relevant to our analysis. Thus, instead of further pre-filtering the dataset, we relied on the LLMs to determine article relevance (see Appendix B).

4 Extraction pipeline and annotations

We focus on two types of response measures to water shortages: water-saving recommendations (*"spara vatten"* in Swedish; i.e., "save water") and water-use restrictions (*"bevattningsförbud"* in Swedish; i.e., "watering bans"). These measures, which signal potential water stress, are communicated to the general public by or on behalf of municipalities and water utilities. We distinguish the measure types to capture the progression from water savings to bans being implemented, continued, expanded spatially, lifted, or avoided (Figure 2). While watering bans require compliance between a clear start and end date (i.e., when imposed and lifted), water savings are more general recommendations that do not follow a strict timeline but may have a measure date.

4.1 Output schema

We extract information on the measure type, dates, location(s), and their cause(s), as shown in Table 1. Each news article may contain information about multiple measures at different locations, but each data entry must correspond to one measure and one location. For each field, in addition to format requirements, constraints may apply, limiting a model's responses to predefined categories or locations within Sweden, or specifying the conditions under which a field should be filled or left blank.

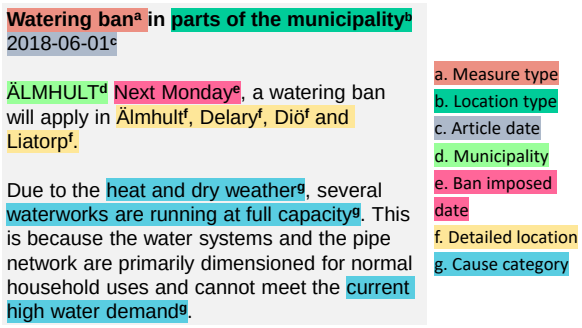


Figure 3: Example article (summarized) highlighting text segments that may inform different extraction fields.

4.2 Gold standard annotation

The annotation process was divided into two steps: first, for article relevance and second, for data extraction (Figure 1). To evaluate how well models could recognize relevant texts, a set of 170 articles was randomly sampled, stratified by year, with at least 2 and up to 15 articles per year. An article is relevant if it discusses a newly introduced, ongoing, or lifted watering ban, or communication that encourages the public to save water due to drought conditions or water supply disruptions. As non-relevant, we consider communication on water supply that does not explicitly encourage the public to take action about water use: debates, opinions, water saving tips, economic or technical innovations, and measures taken at the regional-level that do not directly concern citizens. In the sample, 123 articles (∼70%) were found relevant.

To create a gold standard for data extraction, we expanded the initial sample of 123 relevant articles to 267, representing approximately 2% of the cleaned corpus. The additional articles were also sampled randomly and stratified by year to ensure temporal representation. These articles were randomly divided into two sets: 57 for model development and 210 for testing, with an even representation of dataset years in each set. Annotation of the development set was performed by two annotators to establish clear annotation guidelines, and conflicts were resolved by discussion among the two annotators until consensus was reached on all items. The test set was annotated by a single annotator with the same guideline instructions provided to the LLMs; the second annotator verified cases rated as complex by the first annotator. An example article and its annotations are shown in Figure 3.

Field	Description	Constraints	Format*
article date	Article publication date.	Must be a valid date.	date***
municipality	Municipality concerned by the measure.	LAU 2, corresponding to municipalities in Sweden. Only one municipality may be selected.	string
detailed location	Affected urban areas in a municipality.	Referenced urban area(s) in Sweden. Multiple locations allowed.	string list***
location type	Measure scope.	Categorical: entire municipality or part of the municipality. Only one allowed.	string
measure type	Type of measure applied.	Categorical: ban imposed, ban lifted, pre-emptive measure**, ban continues, ban expanded, or ban avoided. Only one allowed.	string
measure date	Date a measure is announced.	Does not apply to "ban imposed" or "ban lifted" measures.	date***
ban imposed date	Date a ban is imposed.	Only filled if measure type is “ban imposed”.	date***
ban lifted date	Date a ban is lifted.	Only filled if measure type is “ban lifted”.	date***
cause category	Indicates the cause of the measure.	Categorical: lack of rain, high water consumption, warm temperatures, low surface water levels, low groundwater levels, low water levels (unspecified), water repairs, water shortage, power outage, precaution, other, and NA (not available). Must be at least one.	string list

Table 1: Descriptions of extracted fields, including definitions, constraints, and expected formats. (*All dates are in YYYY-MM-DD format. **Preemptive measures refer to water saving recommendations. ***Nullable fields.)

4.3 Prompt development and LLM configuration

The extraction pipeline consists of two main prompts, one for assessing article relevance, returning a binary output (relevant/non-relevant) and the other for data extraction, with separate prompts for each extracted field (Table 1). While all articles are in Swedish, prompts were written in English, and all outputs, except location names, were returned in English, in accordance with the higher performance of LLMs with English-language prompts (Fu et al., 2022; Vela-Tambo et al., 2025). The articles were not translated as Swedish is not usually considered a low-resource language. Although the English language often dominates most open-source and commercial LLM training datasets, the availability of Swedish digital resources and the multilingual capabilities of LLMs suggest that Swedish texts can be processed reliably (Borin et al., 2023). This assumption was tested through model performance evaluation on the relevance and extraction prompts.

For the data extraction prompt, we applied a role-prompting strategy by instructing the model to act as an accurate and precise system, fluent in both Swedish and English. We also used a zero-shot prompting strategy so as not to bias the tested models toward a few examples and to reduce costs. In addition, we relied on explicit variable and category names to reduce ambiguity. Although we aimed for concise prompts, we prioritized comprehensive descriptions over prompt brevity given the relatively large context sizes of all tested models. See Appendix C for a prompt example.

For the relevance classification task, we considered Llama 3.3 (70B) and GPT-5.2[5]. For data extraction, we used LLMs from different model families, including the GPT-5 series (GPT-5.2, GPT-5.4, GPT-5.4-mini, and GPT-5.5) (OpenAI, 2026) and the largest model size of three open-access model families: Llama 3.3 (70B) (Grattafiori et al., 2024), GPT-OSS (120B) (OpenAI et al., 2025), and Deepseek R1 (70B) (Guo et al., 2025).

To benchmark performance, all models received identical prompts. We also set reasoning levels for the GPT-5 models to “none” which improves efficiency and reduces costs as higher reasoning levels require more tokens.[6] For open-access models, we test performance at temperature zero (i.e., deterministic outputs via greedy decoding). While temperature manipulates the token distribution, directly affecting the choice of the next token, reasoning is unique to the GPT-5 series and reflects the

[5]GPT-5.2 was the latest available GPT-5 model at the time.

[6]https://developers.openai.com/api/docs/guides/reasoning

model's capacity to use "internal reasoning" tokens to solve complex tasks.

We performed additional experiments to study the effects of increasing temperature (for open-access models) and reasoning effort (for the GPT-5 series) on data extraction accuracy. We tested the following temperatures: `0.25`, `0.5`, `0.75`, `1.0`, and `1.5`. To account for stochasticity, we averaged error scores of three separate runs. For the GPT-5 models, we run a single experiment for each reasoning level (see Appendix D for details).

The final relevance and data extraction prompts result from an iterative refinement process using GPT-5.2 and GPT-5.5 to maximize performance on the development set. To support prompt development and facilitate error analysis, the models were required to return justifications, text excerpts, and summaries, in addition to the extracted information. Moreover, the models were configured to prioritize output consistency and stability in JSON format. This format also facilitates the evaluation of model outputs against annotations. While the OpenAI API[7] already offers the option for structured JSON output, open-access models do not – therefore, we perform those experiments via the Ollama API.[8]

4.4 Baseline data extraction

To assess the benefits of contextual and semantic understanding in extracting unstructured data, we implemented a rule-based extraction baseline that uses lexical matching and heuristic rules optimized on the development set. Similar to the LLM-based extraction, this deterministic pipeline processes the articles and builds a structured table of measure records, with one entry per measure-location pair.

Location fields were extracted using dictionary matching against predefined localities in Sweden and common lexical variations. The spatial scope of a measure (i.e., the location type, either full or partial coverage of a municipality) is inferred from the presence of sub-municipality locations in the text or specific textual cues. Measure types and their causes were categorized using regex based on keywords and expressions commonly found in Swedish news. Temporal data was extracted using regex and named entity recognition, with relative dates resolved using the article publication date.

We linked measures, locations, and dates based on their proximity in the text. Each measure is associated with a short context window of surrounding sentences, which defines the search scope for locations and dates. While this approach reduces false associations, it may miss information distributed across multiple paragraphs.

5 LLM performance evaluation

For the binary classification task (i.e., categorizing articles as relevant or non-relevant), we used standard evaluation metrics: overall accuracy, precision, recall, and F1 score.

For the data extraction task, distance-based metrics were used to quantify how closely model outputs match the gold annotations. We calculated an aggregated difference metric for each field, representing the distance between the gold standard and the system outputs, ranging from 0 (perfect match) to 1 (complete mismatch), and averaged it over all measures extracted from an article. Missing information was penalized with a "null penalty", setting the error score to 1 for the missing field. These metrics were calculated differently depending on the field data type (Table 2). We supplemented the metrics used by Li et al. (2024) with an additional date-distance metric with a 28-day cutoff, so that the model receives a maximum error score if the date is incorrect by more than 4 weeks. We also derived an overall performance score by aggregating all the field-specific scores.

A single article can yield more than one entry (in our case, 1.6 entries per article on average in the test set), and an LLM may return these measures in any order; therefore, we matched entries found by the LLMs to the gold annotation for each article using three fields: the measure type, the location type, and the municipality. If the LLM misses an entire entry or finds an extra one, it receives the "null penalty" for each field in the extra or missing row, which can significantly increase the error rate. To distinguish errors due to missing or extra entries from errors due to inaccurate outputs, we first calculated a main metric (labeled "All") that accounts for complete mismatch penalization. Then, we repeated the data extraction evaluation by excluding missing or extra entries, thereby eliminating null-penalty effects arising from these additional or omitted measures. This alternative metric is labeled "Match".

[7] https://developers.openai.com/api/docs/guides/structured-outputs

[8] https://docs.ollama.com/capabilities/structured-outputs

Format	Description	Expression
String	Exact location name matching.	$d_{\text{str}}(a, r) := \begin{cases} 0, & \text{if } a = r \\ 1, & \text{otherwise} \end{cases}$ (1)
Dates	Here, Δ_{days} is the absolute difference in days between the dates a and r. If Δ_{days} is longer than 28 days (4 weeks), it is considered incorrect.	$d_{\text{date}}(a, r) := \begin{cases} \frac{\Delta_{\text{days}}}{28}, & \text{if } \Delta_{\text{days}} < 28 \\ 1, & \text{otherwise} \end{cases}$ (2)
Categorical sets	Jaccard distance between two sets of pre-defined categories or location names that must match exactly.	$d_{\text{set}}(a, r) := 1 - \frac{\lvert a \cap r \rvert}{\lvert a \cup r \rvert}$ (3)

Table 2: Difference metrics used for evaluating extracted fields, where a is the annotated value and r is the retrieved value. Difference metrics are in the range $[0, 1]$, where $\downarrow$ lower is better.

5.1 Relatedness classification results

Table 3 outlines recall, precision, accuracy and F1-scores for GPT-5.2 compared to Llama3.3. GPT-5.2 exhibits a slight bias towards over-inclusion (due to producing more false positives). The false positives are articles discussing water saving campaigns and tips to save water; opinions, questions or complaints about water savings; or the effects of water savings on overall consumption. Error analysis shows that while they relate to the overall topic, they do not cover decisions about water-saving measures. The false negatives also comprise borderline cases where messaging about water savings appears, but it is difficult to determine whether it applies to the local- or regional-scale. Llama3.3, on the other hand, produced more false positives and false negatives. The false positives can be traced to the inclusion of context articles on water resources, water supply disruptions, non-municipal measures, or technical and economic initiatives to save water. A third of the false negatives occur due to difficulty in isolating relevant content from irrelevant information, or, in a case specific to Llama3.3, failure to recognize texts as articles, despite prompt instructions that all documents should be treated as such (see Appendix B). However, most false negatives correspond to misclassifying the text, even when the summary of the article content is accurate.

Metric	GPT-5.2 (no reasoning)	Llama3.3:70b (T=0)
Accuracy	0.91	0.60
Recall	0.96	0.72
Precision	0.91	0.73
F1 Score	0.94	0.72

Table 3: LLM performance on relevance classification.

5.2 Data extraction results

Table 4 presents the results of the data extraction by field for the test dataset (along with average scores, coverage, runtime, and costs) for GPT-5.5, the three open-access models, and the baseline extraction (see Appendix E, Table 7 for development set results). Similar average scores on the development and test sets indicate a stable performance for all model families, suggesting that the prompts are sufficiently robust for this use-case. The results that account for all entries (columns "All" in Table 4) show that GPT-5.5 outperforms all open-access models by over 0.3 points. The three open-access models present similar average scores, but for many variables Llama3.3 and Deepseek perform better than GPT-OSS, with lower field-specific error rates. The extraction results are, on average and for several fields, better with the baseline NLP than with open-access models. Specifically, all open-access models show error rates above the baseline error rate for detailed locations, measure type, and cause category, showing that generative AI does not automatically offer advantages for complex extraction tasks. The variables with the greatest error rates are measure type, measure date, and cause category for GPT-5.5. The open-access models struggle with almost all variables except dates of lifted bans.

When considering only matched pairs of entries (error scores in the "Match" columns in Table 4), we see lower error rates, suggesting that missed or extra entries largely influence the overall score. The baseline and all models find fewer entries on average per article compared to the gold-standard test annotations, except for GPT-OSS, which returns more entries. Note that models are permitted to return no entries if they fail to find any relevant

Metric / Set	GPT-5.5		Llama3.3 (70B)		Deepseek R1 (70B)		GPT-OSS (120B)		Baseline	
	All	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**	**All**	**Match**
article date	**0.13**	0.00	0.33	0.01	0.36	0.01	0.55	0.04	0.35	0.00
municipality	**0.17**	0.04	0.52	0.29	0.58	0.34	0.73	0.42	0.69	0.52
detailed location	**0.19**	0.07	0.88	0.82	0.86	0.78	0.92	0.84	0.67	0.50
location type	**0.16**	0.03	0.61	0.42	0.60	0.37	0.77	0.51	0.75	0.62
measure type	**0.27**	0.16	0.76	0.65	0.70	0.53	0.86	0.71	0.69	0.53
measure date	**0.30**	0.26	0.54	0.54	0.53	0.56	0.50	0.61	0.35	0.32
ban imposed date	**0.11**	0.10	0.44	0.53	0.40	0.52	0.40	0.47	0.30	0.34
ban lifted date	**0.03**	0.02	0.09	0.11	0.11	0.14	0.07	0.09	0.11	0.13
cause category	**0.32**	0.22	0.88	0.83	0.80	0.68	0.89	0.76	0.70	0.54
avg. score	**0.20**	0.10	0.57	0.47	0.55	0.44	0.64	0.49	0.51	0.39
avg. score **(dev)**	0.22	0.10	0.56	0.47	0.53	0.43	0.62	0.50	0.54	0.45
coverage	98.57%		100%		100%		100%		100%	
avg. entries per article	1.47		1.18		1.01		1.9		1.25	
runtime (min)	60		62		63		42		0.5	
approximate cost (USD)	10		4		4		5		0	

Table 4: Comparison of extraction performance across open-access models and GPT-5.5 on the **test set**, for both "All" entries versus for an alternative "Match" test set that excludes superfluous or missing entries. Temperatures and reasoning levels are set to 0.0 and "none", respectively. Coverage is the percentage of articles for which the models produced at least one entry. The average score on the development set (dev) is also provided for comparison with the test-set score. The lowest error rates are shown in **boldface**.

information in an article. This only occurs for a few articles with GPT-5.5.

Upon inspecting the outputs, we find that GPT-5 errors can be attributed to prompt interpretation. Date errors often stem from the extent to which the model or the annotators extrapolated dates from the article's date, as authorized by the prompts. In terms of measure type, a frequent error is treating a continuing watering ban as a newly imposed ban, which can be explained by vague formulations (e.g., "there is a ban") in some of the articles. Differences in the detailed locations arise because the model extracts locations such as "west of the municipality", whereas annotators focused only on named locations.

The cause category field is problematic for all GPT-5 models, which can be attributed to its high subjectivity. While most causes are well captured, the lower performance is due to frequent additions to or omissions from the cause category. For example, a common error is the omission of the category "lack of rain", broadly described as dry conditions in the prompts. This may reflect a relatively inclusive annotation approach, which consistently classified references to dry summers as "lack of rain". The model otherwise over-reports the categories "water shortage", "high water consumption", and "other".

Since the cause category is a short multi-label field, using Jaccard distance emphasizes the error score arising from mismatches in the number of reported categories. For this reason, we also include precision, recall, and F1 scores for each cause label in Table 5, for both GPT-5.5 and Llama 3.3. The scores confirm that GPT-5.5 is relatively strong at predicting most labels, distinguishing explicit labels (e.g., "power outage", "water repairs") and fine-grained cause categories (e.g., "low groundwater levels", "water shortage"). However, the model performs poorly for broad categories such as "precaution" and "other" which are not as clearly associated with linguistic indicators. (The same scores for the development set are reported in Table 8, Appendix E.)

For the open-access models, performance is lower across all fields than with GPT-5.5, except for the ban-lifted date. This cannot be explained by interpretation issues. Indeed, the models have difficulty assigning event dates to the correct columns (e.g., measure date versus ban date) and sometimes fail to convert textual relative dates to the desired YYYY-MM-DD format. Open-access models also struggle with categorical variables (i.e., measure types, location types, and causes), which require mapping textual data to predefined labels. For the cause field, for instance, LLama 3.3 defaults to the single label "water shortage" for close to 70% of the returned entries and substantially under-reports labels, predicting only 0.73 labels per entry on average for the test set, compared to 1.71 in the gold

		GPT-5.5			Llama 3.3			
	Label	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
0	High water consumption	0.779	0.918	0.843	0.714	0.068	0.125	73
1	Lack of rain	0.833	0.650	0.730	0.000	0.000	0.000	100
2	Low groundwater levels	0.846	0.819	0.832	0.875	0.149	0.255	94
3	Low surface water levels	0.692	0.643	0.667	0.000	0.000	0.000	14
4	Low water levels unspecified	0.550	0.688	0.611	0.286	0.125	0.174	16
5	NA	0.868	0.575	0.692	0.250	0.012	0.024	80
6	Other	0.200	0.750	0.316	0.000	0.000	0.000	4
7	Power outage	1.000	1.000	1.000	0.800	1.000	0.889	4
8	Precaution	0.474	0.474	0.474	0.000	0.000	0.000	19
9	Warm temperature	0.870	0.839	0.855	0.000	0.000	0.000	56
10	Water shortage	0.667	0.784	0.721	0.161	0.588	0.253	51
11	Water repairs	0.909	0.882	0.896	0.917	0.324	0.478	34
12	Waterworks at capacity	0.816	0.851	0.833	0.000	0.000	0.000	47

Table 5: Precision, recall, and F1 scores for multi-label classification of the cause category for the **test set**. The support column represents the number of examples available in the gold annotations for each label.

annotations and 1.69 predicted by GPT-5.5. As a result, evaluation scores are very poor for most cause categories (Table 5). In addition, when several municipalities are mentioned in a text, open-access models sometimes report them as a single entry rather than as separate entries, as instructed. Finally, the outputs tend to mix Swedish and English, despite instructions to respond in English. For these reasons, the baseline extraction method, which enforces a structured output format and relies on clearly defined rules, achieves a lower overall error rate.

To provide further insights into the identified error types, examples of text excerpts, gold-standard annotations, and model extractions are shown in Tables 11 (for GPT-5.5) and 12 (for Llama 3.3) in Appendix F.

6 Discussion and conclusion

This study benchmarks proprietary LLM and open-access LLM performances against baseline NLP methods for the extraction of complex relational data on a specialized topic: the communication of measures on water use during drought and water supply disruptions. The motivation for the study was to assess the extraction performance without extensive prompt refinement or post-processing, evaluating models under comparable lightly engineered conditions.

As discussed in similar studies (Li et al., 2024; Castro et al., 2024), the performance of proprietary models surpasses that of open-access models, including larger ones, in terms of stability and capacity to generate structured outputs. The limited effects of model temperature and reasoning levels confirm that model size, rather than configuration, is a key factor determining performance.

We find that prompt-engineering alone does not produce better outputs compared to hand-crafted (hence time-consuming) NLP approaches, even for the largest open-access models available from three main LLM families. In this study, prompt optimization may have benefited GPT-5 models, since we did not revise the prompts for the open-access models. However, we suppose that the experimental setup with limited fine-tuning, training, and post-processing explains the lower performance of the open-access models. Specifically, these multilingual models may not have been trained on sufficient amounts of Swedish text. Therefore, fine-tuning on Swedish news articles, more refined guidelines for inferring the cause category, and post-processing could improve the quality of the outputs. A practical implication is that although their access is free, a considerable time investment may be required to achieve satisfactory performance with general-purpose open-access models, creating a trade-off among time, costs, and environmental impacts. To extract structured, precise, and accurate data in this setup, and without any fine-tuning, we have to rely on proprietary models from providers like OpenAI, which is costly. In addition, the extraction process may be difficult to replicate due to model deprecation and the need to refine the prompts for each new model release.

Limitations

We recognize limitations in the development of the gold standard, as inter-annotator agreement was not measured. For the development set, two annotators collaborated to iteratively refine the annotations through joint discussions, which leaves room for bias. For the test set, annotations were performed by a single annotator, while the second annotator reviewed edge cases and proposed revisions where necessary. In addition, both annotators speak Swedish as a second language. This may influence the interpretation of certain journalistic expressions.

The study also presents limitations associated with the experimental setup. First, the prompts were developed for GPT-5.2 and 5.5 and were not revised for other LLMs, which favors the larger models. In addition, limited post-processing and no fine-tuning were performed for the open-access models, steps that would have likely increased model performance. Finally, we were not able to test models extensively trained on Nordic languages due to a low maximum token limit (i.e., context length) and instead relied on multilingual models predominantly trained on English (and Chinese in the case of DeepSeek). Thus, further testing is needed before concluding that open-access models are not suited to complex data extraction for a niche topic.

When applied to the entire dataset, the outputs of the proposed extraction method will be affected by the extent of media coverage of water-use restrictions and the quality of the reported information. The Retriever database includes local news sources, which ensures broad coverage of localized events across Sweden. However, the lack of reporting on a municipality cannot be equated to the absence of a watering ban. Therefore, triangulation approaches combining news media extraction and surveys could also be explored.

Ethics statement

The paper does not raise any ethical issues. The news articles were accessed via a media archive made available through university library agreements, and, in accordance with the copyright and archive data use requirements, no texts were shared outside the research group.

Acknowledgments

J.F. and J.M. acknowledge funding from FORMAS under the project 'Leap in Swedish freshwater monitoring for water supply risk mitigation' (LEAP4WATER, project ID:2022-02148_Formas). J.F. further acknowledges support from CNDS for collaboration with the UFZ. S.Z. acknowledges the ICE Datacenter at RISE Research Institutes of Sweden for providing resources to conduct experiments on open-access models, as well as for providing OpenAI Platform credits to use the OpenAI API. M.M.d.B. discloses funding from the Volkswagen Foundation through the CASCADE project, funded under the Earth Systems Sciences initiative, as well as the "Quantification of the effects of adaptation measures on spatiotemporal changes in flood and drought risk (EXTREME-ADAPT)", an InnoPool project funded by the Helmholtz Association.

References

Lars Borin, Rickard Domeij, Jens Edlund, and Markus Forsberg. 2023. *Language Report Swedish*, pages 219–222. Springer International Publishing, Cham.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are Few-Shot learners.

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.

Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. 2024. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, 82:102742.

Felicia Chiang, Omid Mazdiyasni, and Amir AghaKouchak. 2021. Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature communications*, 12(1):2754.

Mariana Madruga de Brito, Brielen Madureira, Taís Maria Nunes Carvalho, Damien Delforge, Aglaé Jézéquel, Murathan Kurfalı, Ni Li, Gabriele Messori, Joakim Nivre, Barbara Pernici, Niko Speybroeck, Stefano Terzi, Wim Thiery, Bram Valkenborg, Jingxian Wang, Shorouq Zahra, Jakob Zscheischler, and Jan Sodoge. 2026. Assessing socio-economic climate impacts from text data. *Preprint*, arXiv:2605.20793.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask prompt training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Eunmi Lee, Seunghui Choi, Sangeun Lee, Newsha K. Ajami, Hyeoncheol Yoon, Xiaogang He, and Jonghun Kam. 2026. Spatiotemporal and economic impacts of media on water conservation during drought: An explainable ai approach. *Water Resources Research*, 62(6).

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.

Andreas Nicolaidis Lindqvist, Rickard Fornell, Thomas Prade, Linda Tufvesson, Sammar Khalil, and Birgit Kopainsky. 2021. Human-water dynamics and their role for seasonal water scarcity – a case study. *Water Resour. Manage.*, 35(10):3043–3061.

Brielen Madureira, Mariana Madruga de Brito, and Andreas Niekler. 2026. Retrieving floods without floodlights: Topic models as binary classifiers for extreme climate events in german news. *arXiv preprint arXiv:2605.03450*.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2026. Gpt-5.5 system card - openai deployment safety hub.

Kimberly J Quesnel and Newsha K Ajami. 2017. Changes in water consumption linked to heavy news media coverage of extreme climatic events. *Sci. Adv.*, 3(10):e1700784.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Data Mining*, page 1–17. Cambridge University Press.

Michele Ronco, Luca Bandelli, Lorenzo Bertolini, Sergio Consoli, Damien Delforge, Alessio Spadaro, Marco Verile, and Christina Corbane. 2026. Disaster storylines and knowledge graphs from global news with large language models and Retrieval-Augmented generation. *Sci. Data*, 13(1).

Jan Sodoge, 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 Clim. Extrem.*, 41(100574):100574.

Jan Sodoge, Christian Kuhlicke, Giuliano Di Baldassarre, Jan H Fleckenstein, Pia Ebeling, and Mariana Madruga de Brito. 2026. Emerging hotspots of groundwater conflicts during droughts in germany. *iScience*, 29(7).

Statistikmyndigheten SCB. 2023. Kvalitetsdeklaration tätorter.

Claudia Teutschbein, Elise Jonsson, Andrijana Todorović, Faranak Tootoonchi, Elin Stenfors, and Thomas Grabs. 2023. Future drought propagation through the water-energy-food-ecosystem nexus – a nordic perspective. *Journal of Hydrology*, 617:128963.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Javier Vela-Tambo, Jorge Gracia, and Fernando Dominguez-Castro. 2025. CienaLLM: Generative Climate-Impact Extraction from News Articles with Autoregressive LLMs.

Jingxian Wang, Andrea Castelletti, Mariana Madruga De Brito, and Barbara Pernici. 2025. Drought perceived impacts via text mining of social media. *Environmental Research: Water*, 1(4):045007.

Marthe Linda Kris Wens, Michael Hagenlocher, Anastasiya Shyrokaya, Micha Werner, Andrea Toreti, Pedro Henrique Lima Alencar, Amobichukwu Chukwudi Amanambu, Jesse Eugene Bell, Won-Ho Nam, Sumiran Rastogi, Juan Rivera, Celia Ramos Sánchez, Matteo Zampieri, Anne-Sophie Sabino-Siemons, Kelly Smith, Faranak Tootoonchi, Patricia Trambauer, Michelle Theresia Helena van Vliet, Marit van Thiel, and 5 others. 2026. Ten key insights and gaps to inform drought risk research, policy and practice. *Nat. Water*, 4(6):711–721.

A Further details about the corpus development

Table 6 presents the keyword searches applied to select the articles in the Retriever media archive. To assess the relevance of the articles retrieved for each query, we examined the titles and headlines of the top 20 most relevant and the 20 least relevant articles. If a majority of articles were relevant in this subset, we included the full query in our search.

To filter out texts that did not mention municipality or urban area names, we used a list of official Swedish municipalities and the most recent list of urban areas from Statistics Sweden[9]. Urban areas are updated yearly, but we assume that year-to-year variations are negligible. Using regex patterns, the article's main text was searched for municipality names in their nominative and genitive forms, including demonym suffixes. Urban areas were searched for only in their nominative form.

Figure 4 shows the distribution of cosine similarity values computed from TF-IDF representations of the articles after filtering out texts that did not contain municipality or urban area names. Most article pairs have low to moderate similarity (below 0.7), but there is a visible cluster of highly similar texts above 0.8–0.9. This distribution and the manual examination of text clusters confirmed that 0.8 can be used as a threshold value to identify duplicate and near-duplicate articles.

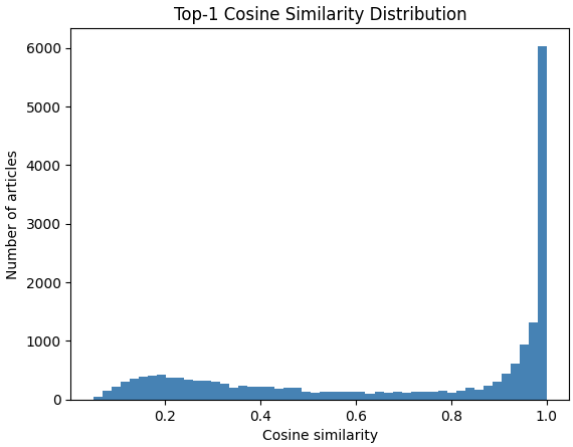


Figure 4: Cosine similarity distribution

B Relevance classification prompt

Prompt 1: Relatedness classification

You will be given a news article in Swedish. Determine whether the article is relevant. An article is considered relevant if it contains information about:

- Water use restrictions or irrigation bans (e.g. bevattningsförbud) issued by a municipality or a water utility
- Official water saving recommendations
- Statements that no restrictions are needed

This information typically concerns municipal water management and often comes from a municipality or municipal water utility (VA-bolag), but the issuing authority may not always be explicitly stated in the text.

An article is NOT relevant if it only discusses:

- Water abstraction bans or permit restrictions issued by a regional authority (länsstyrelsen)
- Hypothetical bans that have not been formally decided
- Water saving technologies or tips for households
- General advice on saving water without official recommendations
- Economic or political discussions about water
- Water pricing or taxation
- Environmental issues without specific restrictions or recommendations

Even short articles can be relevant. If municipal restrictions or recommendations are mentioned, even briefly, the article is still considered relevant if it provides concrete information such as the location, timing, or implementation of the measure. Return True if the article is relevant. Return False if the article is not relevant.

REMEMBER: All texts provided are articles, even if they look like briefs, summaries, metadata, article lists. Ignore non-relevant content, only focus on relevant content.

[9] https://www.scb.se/hitta-statistik/statistik-efter-amne/boende-bebyggelse-och-mark/bebyggelseomraden/tatorter-och-smaorter

	Target	Query	Articles
1	Water restrictions	`bevattningsförbud* OR bevattningsstopp OR ((förbud* OR förbj* stopp* OR begräns ) NEAR/2 (bevattn* OR vattna)) OR vattenrestriktion* OR bevattningsrestriktion* OR vattningsrestriktion* OR (restriktion* NEAR/2 vattenanvändning*) OR (totalförbud near/2 vatten)`	12154
2	Save water	`(spara OR bespara OR sparsam*) NEAR/3 (vatten OR vattnet OR dricksvatten OR dricksvattnet) OR vattenbesparing`	6672
3	Save water (synonyms and expressions)	`((minska OR ''dra ner'' OR ''skära ner'' OR begräns* OR anpassa OR ''håll ner'' OR ''använda mindre'' OR försiktig* OR sparsam*) NEAR/2 (vattenförbrukning* OR vattenanvänd*)) OR ((hushålla OR ''inte slösa'' OR ''slösa inte'' OR ransonera OR snåla OR ''tappa upp'') NEAR/2 (vatten OR vattnet OR dricksvatten OR dricksvattnet)) OR ((uppmana* OR manar OR varnar OR varna OR vädja*) NEAR/5 (bevattn* OR vattenförbrukning*))`	3802
4	Water shortages	`kommun* NEAR/5 (vattenbrist OR ''brist på vatten'' OR ''dricksvattenbrist'')`	976

Table 6: Search queries used to construct the document corpus and total articles retrieved on 25/02/2026

C Data extraction prompt example

Prompt 2: Extract the municipality

Identify the Swedish municipality or municipalities (kommun/kommuner) where the extracted measure_type applies. An article may describe multiple measures and multiple municipalities. For each municipality, create a separate output entry associated with the relevant measure_type.

- If the same measure_type applies to multiple municipalities, return one entry per municipality (do NOT group municipalities together).
- If different municipalities are associated with different measure_types, assign each municipality to the correct measure_type and return separate entries.

Important:

- The municipalities must be official Swedish municipalities (kommuner).
- Only include municipalities that are explicitly mentioned in the text verbatim and directly linked to the measure.
- When returning municipality names, output only the municipality name itself (e.g. "Uppsala", even if the text says "Uppsala kommun", or "Region Uppsala").
- Do not normalize or translate municipality names.
- You may only infer the municipality if it can be identified with high certainty from the article source (e.g. "Uppsala kommun", "24 Kalmar") or context (i.e., municipality inhabitants).
- Do not include municipalities that are not affected by, subject to, or clearly linked to a water measure.
- If there is any ambiguity, do not return a municipality.

D Deterministic output for OpenAI models

Even though only low or no reasoning levels are recommended by OpenAI for deterministic workflows (https://developers.openai.com/cookbook/examples/gpt-5/gpt-5_new_params_and_tools#42-takeaways), OpenAI also uses prompt caching for the GPT-5 series, among others, making it difficult to determine if multiple runs are producing identical results or simply reliant on an internal caching mechanism to reduce latency (https://developers.openai.com/api/docs/guides/prompt-caching).

E Additional Results

This section contains additional results, including development set results for the preliminary experiment in Table 7, additional evaluation scores for the cause category classification on the development set in Table 8, comparison of performance on the test-set across different GPT-5 series models in Table 9, and the performance on the test set across ascending temperatures for the Llama3.3 model compared to GPT-5.5 performance across different reasoning levels in Table 10.

F Examples of errors by GPT-5.5 and Llama 3.3

This section presents article text excerpts translated into English, together with the corresponding annotations and selected examples of model errors for different extraction fields. The examples of errors are presented in Table 11 (GPT-5.5) and Table 12 (Llama 3.3).

Metric / Set	GPT-5.5 All	GPT-5.5 Match	Llama3.3 (70B) All	Llama3.3 (70B) Match	Deepseek R1 (70B) All	Deepseek R1 (70B) Match	GPT-OSS (120B) All	GPT-OSS (120B) Match	Baseline All	Baseline Match
article date	**0.18**	0.00	0.31	0.02	0.31	0.00	0.45	0.00	0.30	0.02
municipality	**0.24**	0.07	0.49	0.27	0.47	0.23	0.61	0.30	0.67	0.54
detailed location	**0.26**	0.10	0.87	0.82	0.87	0.81	0.91	0.84	0.72	0.60
location type	**0.25**	0.08	0.60	0.43	0.54	0.34	0.80	0.63	0.79	0.70
measure type	**0.26**	0.10	0.72	0.60	0.64	0.48	0.84	0.71	0.81	0.73
measure date	**0.35**	0.27	0.64	0.62	0.62	0.61	0.54	0.62	0.36	0.24
ban imposed date	**0.06**	0.04	0.48	0.54	0.40	0.54	0.48	0.51	0.25	0.30
ban lifted date	**0.02**	0.03	0.04	0.06	0.10	0.14	0.04	0.06	0.05	0.06
cause category	**0.34**	0.20	0.88	0.85	0.82	0.74	0.90	0.82	0.88	0.84
avg. score	**0.22**	0.10	0.56	0.47	0.53	0.43	0.62	0.50	0.54	0.45
coverage	96.42%		100%		100%		96.43%		100%	

Table 7: Comparison of extraction performance on the **development set**. Temperatures and reasoning levels are set to 0.0 and "none", respectively. The results under the "Match" columns shows the result when extra or missed entries found by the LLMs are ignored. Coverage is the percentage of articles the models produced at least one entry for.

		GPT-5.5			Llama 3.3			
	Label	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Support**
0	High water consumption	0.762	0.800	0.780	0.600	0.150	0.240	20
1	Lack of rain	0.806	0.833	0.820	0.000	0.000	0.000	30
2	Low groundwater levels	0.944	0.850	0.895	1.000	0.150	0.261	20
3	Low surface water levels	0.615	0.800	0.696	0.000	0.000	0.000	10
4	Low water levels unspecified	——	——	——	——	——	——	0
5	NA	0.833	0.714	0.769	0.000	0.000	0.000	21
6	Other	0.500	0.500	0.500	0.000	0.000	0.000	4
7	Power outage	——	——	——	——	——	——	0
8	Precaution	0.000	0.000	0.000	0.000	0.000	0.000	2
9	Warm temperature	1.000	1.000	1.000	0.000	0.000	0.000	25
10	Water repairs	1.000	1.000	1.000	0.500	0.167	0.250	6
11	Water shortage	0.519	0.824	0.636	0.222	0.706	0.338	17
12	Waterworks at capacity	0.800	0.889	0.842	0.000	0.000	0.000	9

Table 8: Precision, recall, and F1 scores for multi-label classification of the cause category for the **development set**. The support column represents the number of examples available in the gold annotations for each label. Labels with rows marked as dashes ('——') do not appear in the development set gold annotations.

Metric / Set	GPT-5.5 All	GPT-5.5 Match	GPT-5.4 All	GPT-5.4 Match	GPT-5.2 All	GPT-5.2 Match	GPT-5.4-mini All	GPT-5.4-mini Match
article date	**0.13**	0.00	0.15	0.00	0.14	0.00	0.22	0.00
municipality	**0.17**	0.04	0.21	0.08	0.22	0.09	0.28	0.07
detailed location	**0.19**	0.07	0.22	0.09	0.23	0.11	0.32	0.13
location type	**0.16**	0.03	0.20	0.07	0.19	0.06	0.27	0.06
measure type	0.27	0.16	**0.25**	0.13	0.28	0.16	0.36	0.18
measure date	0.30	0.26	**0.24**	0.19	0.30	0.29	0.43	0.38
ban imposed date	**0.11**	0.10	0.13	0.12	0.14	0.13	0.16	0.16
ban lifted date	**0.03**	0.02	0.04	0.02	0.07	0.06	**0.03**	0.02
cause category	**0.32**	0.22	0.38	0.28	0.37	0.27	0.49	0.35
avg. score	**0.19**	0.10	0.20	0.11	0.21	0.13	0.28	0.15
coverage	91.56%		91.57%		92.77%		81.33%	

Table 9: Comparison of extraction performance on the **test set** for different GPT-5 models at reasoning level "none". The results under the "Match" columns shows the result when extra or missed entries found by the LLMs are ignored. Coverage is the percentage of articles for which the models produced at least one entry. The lowest error rates are shown in **boldface**.

	Llama 3.3 (70B) temperature						GPT-5.5 reasoning level			
set	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**1.5**	**none**	**low**	**med.**	**high**
article date	0.31	0.33±0.00	0.36±0.02	0.37±0.02	0.40±0.02	0.43±0.01	0.13	0.13	0.15	0.13
municipality	0.49	0.53±0.01	0.58±0.04	0.57±0.01	0.60±0.02	0.62±0.01	0.17	0.17	0.19	0.16
detailed location	0.87	0.87±0.00	0.87±0.03	0.84±0.03	0.83±0.03	0.85±0.02	0.19	0.19	0.20	0.17
location type	0.60	0.60±0.01	0.60±0.03	0.60±0.01	0.64±0.01	0.63±0.02	0.16	0.16	0.18	0.16
measure type	0.72	0.75±0.02	0.78±0.01	0.76±0.02	0.80±0.02	0.80±0.01	0.27	0.24	0.26	0.27
measure date	0.64	0.54±0.03	0.55±0.02	0.54±0.03	0.51±0.01	0.52±0.04	0.30	0.30	0.30	0.25
ban imposed date	0.48	0.43±0.01	0.41±0.02	0.37±0.02	0.35±0.02	0.34±0.03	0.11	0.14	0.14	0.14
ban lifted date	0.04	0.10±0.02	0.12±0.01	0.10±0.01	0.12±0.02	0.12±0.01	0.03	0.03	0.02	0.03
cause category	0.88	0.91±0.00	0.90±0.01	0.89±0.01	0.90±0.01	0.91±0.02	0.32	0.35	0.36	0.33
average score	0.56	0.56±0.00	0.57±0.01	0.56±0.01	0.57±0.00	0.58±0.01	0.19	0.19	0.20	0.18

Table 10: Performance on all entries for Llama3.3 at different temperatures and GPT-5.5 at different reasoning levels on the **test set**. Evaluation is performed on all articles. Standard deviation is shown for Llama3.3 for non-deterministic runs with temperature > 0, representing the variance between three separate runs. GPT-5 runs at different reasoning levels were performed only once.

Error	**Text excerpt (translated)**	**True**	**Predicted**
Measure type	"In several municipalities, a watering ban or a partial watering ban has already been imposed. For example Flen, [...]"	ban continues	ban imposed
Measure date	"Date: 2020-06-17", "Since earlier this year there has been a ban in Båstad and Osby municipalities"	Not mentioned	2020-06-17
Cause category	"[...] avoid using a hose for watering and please be mindful of your consumption.", "While water supplies in Bollnäs and Ovanåker remain plentiful for now, residents are still urged to save water." "That recommendation was introduced [...], when temperatures began to rise..."	Warm temperature, Precaution	Warm temperature, High water consumption
Cause category	"Water scarcity is spreading across the country, and Karlshamn is no exception. Alarmingly low water levels in Lake Mien [...].", "... the precipitation [this morning] could not have come at a better time."	Lack of rain, Low surface water levels, Water shortage	Low surface water levels, Water shortage
Number of entries	"At present, there are bans in 12 municipalities including Östhammer in Uppsala county.", "In Enköping we have low water levels and encourage everyone to save water."	2 separate entries: Uppsala, Enköping	Only 1 entry reported: Enköping

Table 11: Examples of errors by GPT-5.5 (at reasoning effort 'none')

Error	Text excerpt (translated)	True	Predicted
Municipality and number of entries	"There is a ban in Högsby and Hultsfred municipalities [...]"	"Högsby" (one entry), "Hultsfred" (second entry)	"Högsby and Hultsfred" (single entry)
Detailed location	"The water levels in our groundwater aquifers are still low..."	"Not mentioned"	"our groundwater aquifers" (returned in Swedish)
Location type	"Gotland", "Acute water shortage [...] in Visby"	Part of the municipality	Entire municipality
Measure date	"Ban in Flöghult municipality during the summer"	"Not mentioned"	"2021 summer"
Ban imposed date	"Date: 2019-06-13 [...] The ban is from July 4th to August 6th [...]"	2019-07-04	2019-06-13
Ban lifted date	"the ban remains until further notice"	"Not mentioned"	"until further notice" (returned in Swedish)
Cause category	"[...] the cause of the watering ban is the low groundwater levels".	Low groundwater levels	Water shortage

Table 12: Examples of errors by Llama 3.3 (using deterministic output)

Ontology-Guided Query Expansion for Evidence Retrieval in ESRS E1 Climate Disclosures

Tatsuya Ishisaka
ASUENE Inc.
ishisaka.tatsuya@asuene.com

Abstract

Auditable analysis of corporate climate disclosures requires *retrieving*, for each regulatory disclosure requirement, the report page or section that contains the evidence. We study **ontology-aware query expansion** (OAE) for **ESRS E1** (Climate Change) requirements over a new page/section-level benchmark: **five companies**, 1,177 real-artifact chunks, and 114 human-reviewed gold labels, plus a held-out ESRS-native company (64 labels) unseen during ontology construction. Against lexical, dense, hybrid, and pseudo-relevance-feedback (PRF) baselines, a lightweight ontology built from the standard's requirements improves retrieval: the best configuration reaches Recall@10 **0.842** vs. BM25's 0.702 (McNemar p=.003). Generic PRF expansion *hurts* retrieval while ontology-guided expansion helps; a corpus-blind LLM query expander is a far stronger control and ties with the ontology in-domain, but does not transfer to the held-out company, where OAE stays ahead of lexical, dense, and LLM-expansion baselines while being statistically indistinguishable from cross-encoder reranking — a comparison we treat as *diagnostic* rather than conclusive. We provide an error analysis of the residual misses. "Gold" denotes retrieval ground truth only; we make no company scoring or compliance judgement.

1 Introduction

Under the EU Corporate Sustainability Reporting Directive, companies report against the European Sustainability Reporting Standards; ESRS E1 (Climate Change) defines concrete disclosure requirements — transition plans, climate targets with base and target years, energy consumption and renewable shares, and gross GHG emissions by scope (European Financial Reporting Advisory Group (EFRAG) and European Commission, 2023). ESRS E1 is, in effect, a checklist of disclosure requirements a company must address.

Company sustainability reports, however, are free-form PDFs of 100–200 pages, and their layout, structure, and level of detail vary by company. This holds even for ESRS-native reports, since the standard does not prescribe where each item must appear or under which heading, so where a given checklist item is stated still varies by company. In addition, the regulatory wording of the checklist and the wording a company actually uses in its report often do not match, so using the regulatory wording directly as a search query does not reliably find the relevant passage. The problem therefore combines two factors: free-form layout and vocabulary mismatch.

For analysts, auditors, and regulators, accurately and quickly locating the passage that addresses each checklist item is essential. Constructing such a *disclosure matrix* (requirement $\times$ company, populated with evidence pointers) thus begins with a retrieval problem: rank a company's report pages/sections so the evidence surfaces at the top for human verification.

This vocabulary and structural mismatch makes the retrieval problem harder than generic ad-hoc search: sustainability reports are saturated with climate-adjacent boilerplate, which makes topically "close" pages adversarial distractors. Classical pseudo-relevance feedback, which expands the query with terms from top-ranked documents (Lavrenko and Croft, 2001; Carpineto and Romano, 2012), is prone to *query drift* in exactly this setting. We instead expand queries from a *disclosure ontology*: requirement labels, curated evidence-query phrases, field names, synonyms, units, and cross-standard mappings.

RQ1. *Does ontology-aware query expansion improve evidence retrieval for disclosure requirements? Is the benefit attributable to the ontology's curated content, rather than to query expansion in general, to score fusion, or to what a general-*

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 116–123
October 28, 2026 ©2026 Association for Computational Linguistics

purpose LLM would propose unprompted by the standard? And how does query-side expansion compare with rerank-stage alternatives?

Contributions. (1) A **human-reviewed page/section-level ESRS E1 evidence-retrieval benchmark**: 5 companies, 12 requirements, 114 gold labels over 1,177 real-artifact chunks, each subset checked by an independent blinded verification pass (κ=0.67 and 0.94) and protected by an evaluation-contamination guard, plus a fair-pooled **held-out ESRS-native company** (64 labels). (2) **Ontology-aware query expansion** from a lightweight, auditable regulation ontology, with LITE/FIELD/FULL field tiers evaluated across three independent scorers. (3) A **controlled decomposition** of where the gain comes from — matched fusion, PRF, and LLM-expansion controls, held-out transfer, and paired significance tests — showing that curated ontology content, not expansion per se, is the operative factor, and identifying where page-level reranking and unconstrained LLM expansion fall short. We scope this paper to retrieval; extraction-stage validation is future work so that retrieval is established as stable first.

2 Related Work

Climate disclosure retrieval. We build on a recent line studying retrieval over corporate climate disclosures and expert-informed query reformulation, but focus on ESRS E1 requirement-grounded evidence retrieval and isolate the effect of curated ontology content through matched controls. Schimanski et al. (2024) release CLIMRETRIEVE, an expert-annotated IR benchmark over 30 sustainability reports with 16 analyst questions and graded relevance on ∼350-word paragraphs, and show that progressively replacing the raw question with longer expert-informed explanations improves embedding retrieval. Ahmed et al. (2025) build ESG-CID by mining companies' own disclosure content indexes as weak supervision for requirement-to-section retrieval across GRI and ESRS, with an ESRS evaluation split, and show fine-tuned BERT encoders can beat commercial embeddings. Neither displaces the present contribution: relative to CLIMRETRIEVE, our queries are ESRS E1 *requirements* rather than analyst questions and our expansion content is authored from the standard rather than distilled from labelled passages, tested with a budget-matched PRF control, an LLM-expansion control, replication across three scorers, and a held-out ESRS-native company; relative to ESG-CID, our E1 labels are human-reviewed and adjudicated rather than content-index weak supervision, and we isolate *which* ontology-derived query fields cause the gain rather than fine-tuning an encoder.

Query expansion and retrieval baselines. Automatic query expansion spans pseudo-relevance feedback (Lavrenko and Croft, 2001; Carpineto and Romano, 2012) and ontology/knowledge-base methods (Bhogal et al., 2007); we compare both, plus an LLM expander, under an identical scorer. We situate OAE against BM25 (Robertson and Zaragoza, 2009), dense sentence encoders (Reimers and Gurevych, 2019; Karpukhin et al., 2020), rank fusion (Cormack et al., 2009), and BERT cross-encoder reranking (Nogueira and Cho, 2019); full configurations are given in §4.

3 Task and Data

Task. For a disclosure requirement r and a company's reporting corpus, order chunks so the gold evidence chunk(s) rank highly. The 12 ESRS E1 evaluation requirements span transition plan (2), targets (4), energy (2), GHG emissions (3), and gross/removals separation (1).

Corpus (real artifacts). Official reporting artifacts from five companies, parsed to **1,177 chunks** (per-company breakdown in Appendix D): PDF pages and HTML sections spanning US and EU/CSRD reporters (one report is ESRS-native), so requirement phrasing is evaluated against both native and non-native disclosures.

Gold annotation and verification. All gold is **human-reviewed** over ranked candidate chunks per (requirement, document), yielding **114 labels** (92 direct, 22 partial) covering all 12 requirements and all 5 companies. Reliability was measured by *independent blinded* verification passes (given only the requirement, its codebook notes, and the chunk text; no access to the gold). On Companies B–D the blinded pass agreed at 84.3% (Cohen's κ=0.67); all 13 disagreements were adjudicated with document-level verification (5 codebook-violating labels removed, 7 best-available-evidence labels relabelled *partial*). On the two EU reporters, where the human reviewed AI-suggested candidates, a same-protocol pass agreed at 97.4% (κ=0.94). We treat κ=0.67 as

the headline reliability, since the 0.94 figure is likely optimistic (AI-assisted review, and the verifier shares the model family that produced the candidates). A validation guard rejects labels referencing nonexistent chunks, low-confidence silver labels, local-snapshot artifacts, or document-level smoke results. Our **annotation protocol** defines **"gold" as retrieval ground truth only**: the labels record where document-grounded evidence is located, and support no claim about company quality, disclosure adequacy, or compliance.

Held-out company. To check generalisation beyond the five development companies, we add **Company F** (2025 Sustainability Statement, 158 pages, ESRS-native), used in no ontology authoring, tuning, or pooling decision. Candidates were pooled as the union of every compared method's top-k (BM25, Hyb-SBERT, OAE-lite, CE-rerank), blind-labelled, and re-checked for completeness as each method was added — yielding **64 gold labels** (12/12 requirements) with a pool neutral to any one system's query vocabulary.

4 Methods

All methods rank the same 1,177 chunks (restricted to the label's company at evaluation time) over the concatenation of section title and chunk text.

Baselines. BM25: Okapi BM25 (k_1=1.2, b=0.75) on the raw `requirement_text`. LSA: TF–IDF (uni+bigrams, 8k features) reduced by truncated SVD to 64 dimensions, cosine. SBERT: all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych, 2019), cosine. BGE: bge-small-en-v1.5 retrieval embeddings with its query prefix. HYBRID: equal-weight sum of min–max-normalised BM25 and dense scores (HYB-LSA, HYB-SBERT, HYB-BGE). The three standalone dense scorers all underperform their hybrids and are omitted from Table 1 for space. PRF-RM3: relevance-model expansion (Lavrenko and Croft, 2001) — top-20 terms from the top-10 BM25 chunks (weighted by normalised score × relative tf × IDF), scored by the same HYB-LSA scorer as the OAE variants. LLM-QE: an LLM query expander, the natural low-cost competitor to a hand-curated ontology. A fixed prompt (Appendix C) shows a current-generation instruction-tuned LLM *only* the standard name and the `requirement_text` — no corpus, no gold labels, no ontology — and asks for at most 20 expansion terms: exactly PRF-RM3's budget, inside the 13–24 terms OAE-lite adds. The generated expansions were frozen before retrieval evaluation, making all reported runs deterministic. Unlike an LLM *reranker*, an expander never sees the corpus and so does not inherit the circularity risk of our partly AI-assisted labels. We score it with HYB-LSA and HYB-SBERT, matching the two OAE-lite configurations. CE-RERANK: a BERT cross-encoder (Nogueira and Cho, 2019) (`ms-marco-MiniLM-L-6-v2`) reranks the company-restricted top-50 of a seed ranking (CE-RERANK(BM25) or CE-RERANK(HYB-SBERT)) by scoring (`requirement_text`, page text, truncated to 2k chars) pairs; chunks beyond rank 50 keep their seed order. Base retrievers already reach Recall@50 of .99–1.00, so the reranker has near-full candidate coverage.

Ontology construction. We designed the ontology as a lightweight, auditable retrieval artifact rather than a full OWL/SHACL implementation. Each ESRS E1 requirement node stores a label, evidence-query phrases, fields, units/values, and validation notes, from which the LITE/FIELD/FULL query variants are generated. Expansion terms are authored from the standard and general domain vocabulary, never from gold pages; Appendix B details the workflow and leakage guardrails.

Ontology-aware expansion (OAE). The query is the requirement text plus ontology fields: OAE-LITE adds label + curated evidence-query phrases; OAE-FIELD adds required field names; OAE-FULL adds synonyms, units/values, and cross-standard mappings. Expanded queries are scored by HYB-LSA; to test scorer-independence we also score the lite expansion with HYB-SBERT and HYB-BGE.

Metrics and protocol. Recall@5, Recall@10, MRR, and page-level accuracy; primary metric Recall@10 (and MRR for expansion contrasts). Significance: exact McNemar on paired hit@k and sign-permutation (10^5 resamples) on reciprocal ranks; Holm correction within the expansion family; fixed seeds. Only the human+LLM `manual_page_gold` set is used; document-level smoke and snapshot-based results are excluded by a guard.

method	R@5	R@10	MRR	p.acc
Baselines				
BM25	0.456	0.702	0.305	0.140
Hyb-LSA	0.465	0.693	0.296	0.132
Hyb-SBERT	0.491	0.693	0.297	0.123
Hyb-BGE	0.500	0.719	0.287	0.114
Generic expansion / reranking				
PRF-RM3	0.518	0.632	0.305	0.140
LLM-QE	**0.614**	0.807	**0.365**	**0.184**
LLM-QE$_{\text{SBERT}}$	0.596	0.798	0.345	0.167
CE-rerank(BM25)	0.430	0.693	0.276	0.123
CE-rerank(Hyb-SBERT)	0.412	0.658	0.277	0.132
Ontology expansion				
OAE-lite	**0.614**	0.807	0.358	0.149
OAE-lite$_{\text{BGE}}$	**0.614**	0.816	0.353	0.149
OAE-lite$_{\text{SBERT}}$	0.605	**0.842**	0.353	0.149

Table 1: Retrieval on `manual_page_gold` (n=114, 5 companies, human-reviewed). Best per column in bold (ties bolded); additional variants are omitted for space (§4). In-domain, LLM-QE is indistinguishable from OAE-lite (Appendix Table 3); the two separate on the held-out company (Appendix Table 2).

5 Results

Main result. The best system, OAE-LITE$_{\text{SBERT}}$, reaches Recall@10 **0.842** vs. 0.702 for BM25 (McNemar p=.0025; R@5 p=.008); full pairwise p-values are in Appendix A.

Where does the gain come from? Ontology expansion improves over every baseline on all three scorers, while PRF degrades its own scorer. *(i) Not fusion*: Hyb-LSA over BM25 is not significant ($p\geq$.65). *(ii) The expansion itself*: lite expansion is significant on MRR for all three scorers (LSA p=.014, SBERT p=.042, BGE p=.0015, Holm-corrected) and on SBERT Recall@10 (p=.0002, 19 discordant pairs vs. 2). *(iii) Ontology content, not expansion per se*: PRF-RM3 drops Recall@10 to 0.632, below its own unexpanded scorer, and loses to OAE-lite (p=.0002). *(iv) But a corpus-blind LLM expander is a far stronger control*: it clears PRF-RM3 and BM25 (R@10 $p<$.0001/.029) and ties the ontology in-domain (p=1.0/.69), taking the best MRR/page-accuracy in Table 1 — substitutable in-domain, not out of domain (below).

Robustness and ablation. The effect holds regardless of company set (US n=52 p=.002; EU/CSRD n=62 ontology-vs-PRF p=.001); across OAE tiers, lite captures the benefit — field/full do not beat it ($p\geq$.29) and cost more to maintain.

Cross-encoder reranking (diagnostic). We treat this comparison as diagnostic rather than conclusive: both CE-RERANK variants (Table 1) degrade their own seed rankings — misapplied rather than outmatched, likely because the tested model is trained for short passages and truncates long report pages. A page-length-matched reranker is the appropriate comparison and remains future work.

Zero-shot generalisation (held-out company). On the held-out ESRS-native Company F, OAE-lite continues to transfer: it improves over BM25 (R@10 .812 vs. .594, p=.009) and separates from LLM-QE (R@10 .812 vs. .656, p=.041), although the latter comparison is suggestive at n=64 and uncorrected. The cross-encoder closes the in-domain gap (.734/.750 R@10), and OAE-vs-CE is not significant ($p\geq$.75). We therefore claim zero-shot generalisation against lexical, dense, PRF, and LLM-expansion baselines, not superiority over a modern reranker; the full held-out table is in Appendix Table 2.

6 Error Analysis

The remaining errors are concentrated in target-family requirements: 13 of the 18 missed gold labels @10 involve target statements, target years, or base years. Most are ranking inversions among topically correct pages in target-dense reports, where reporting-period and target/base years compete. Appendix E shows which expansions help and hurt: OAE helps when regulatory terms and report headings diverge (gross/removals separation, target year), but can over-focus when multiple pages are valid evidence for one requirement.

7 Conclusion

We introduced an ESRS E1 page/section-level evidence-retrieval benchmark and showed that ontology-aware query expansion improves requirement-grounded retrieval beyond BM25, hybrid retrieval, and generic PRF, with the effect replicated across scorers and company subsets. A corpus-blind LLM expander is a strong in-domain control, but OAE transfers better to a held-out ESRS-native company while remaining statistically tied with the tested cross-encoder reranker. These results support lightweight, auditable regulation ontologies as practical retrieval guardrails; page-length-matched reranking and extraction-stage validation are the next steps.

Limitations

Small benchmark. Our benchmark remains small: five development companies and one held-out ESRS-native company, 12 requirements, 114 + 64 labels, English-language reports. The results should be read as evidence from a focused pilot benchmark, not as broad generalisation across sectors, jurisdictions, and reporting styles. Expanding the company and requirement coverage is the single most useful extension.

Single primary annotator. Gold labels rely primarily on one human reviewer with independent *blinded* verification passes, not full independent double annotation of every item; the verifier is an LLM rather than a second human, so human–human agreement is unmeasured. We report the conservative of our two agreement estimates (κ=0.67) as the headline and describe the annotation protocol in §3 to make the labelling contract explicit.

Retrieval-only evaluation. This work evaluates the retrieval step that precedes extraction and verification. We make no claim of end-to-end improvement in disclosure drafting, assurance, or regulatory analysis, and we do not evaluate whether better-ranked evidence yields better extracted values — that is exactly the next stage of this line of work.

Cross-encoder comparison is diagnostic. The rerankers we test are trained for short passages and truncate long page-level inputs, which is the most likely reason they can degrade their own seed rankings here. This makes them an informative diagnostic about granularity but a poor basis for any claim about neural reranking in general; a reranker matched to page-length inputs (or a passage-level decomposition of each page) is needed for a fair comparison and is left to future work.

LLM query expansion. Our LLM-QE baseline uses one frozen generation per requirement from a single model and a single fixed prompt. We did not sweep prompts, models, sampling temperature, or multi-query expansion, all of which could move it in either direction; its in-domain parity with the ontology should therefore not be read as an upper bound on LLM expansion. Our gold is also partly AI-assisted by a related model family — an expander is far less exposed to this than a reranker, since it never sees the corpus, but the risk is not zero. For the same reason we exclude an LLM-*as-reranker* baseline, and a comparison against direct LLM prompting (asking a model to locate evidence with no retrieval step) remains open; our retrieval-based design is motivated by deterministic, inspectable evidence pointers rather than a demonstrated accuracy or cost advantage over LLM-direct-answering, which independently annotated labels would be needed to evaluate fairly.

Metric and tuning caveats. Page-level accuracy is low for all methods ($\leq$ 0.19): adjacent pages share vocabulary, so exact-page top-1 is harsh — we treat Recall@10 and MRR as primary. The main benchmark's candidate-pool seed vocabulary overlaps more with OAE's evidence-query terms than with BM25's (Jaccard 0.87 vs. 0.79), so the absolute margin on this benchmark (Table 1) may be somewhat optimistic; the fair-pooled Company F held-out check (Table 2, §5), which is neutral to any one system's query vocabulary, shows the qualitative conclusion is unaffected — if anything, the margin there is larger. That check is itself limited to one company (n=64), so the comparisons made on it — OAE vs. CE-reranker, and OAE vs. LLM-QE — are underpowered (§5). We did not tune fusion weights, PRF depth, LLM-QE budget, or BM25/expansion hyperparameters; all comparisons share untuned defaults.

Ethics and evaluation safeguards. We analyse only publicly released sustainability reports and make **no company-scoring, compliance, weakness-detection, or "not disclosed" judgement**; "gold" denotes retrieval ground truth only. We refer to companies by anonymized labels to avoid the appearance of assessing specific companies' disclosure quality. Randomised procedures use fixed seeds, and the LLM-QE generations were frozen before evaluation. Paper results use only the `manual_page_gold` set; snapshot-based and document-level smoke results are excluded.

References

Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. 2025. Enhancing retrieval for ESGLLM via ESG-CID: A disclosure content index finetuning dataset for mapping GRI and ESRS. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 1–16, Vienna, Austria. Association for Computational Linguistics.

Jagdev Bhogal, Andrew Macfarlane, and Peter W. H. Smith. 2007. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 758–759. ACM.

European Financial Reporting Advisory Group (EFRAG) and European Commission. 2023. ESRS E1 climate change — european sustainability reporting standards, annex i to commission delegated regulation (eu) 2023/2772. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg_del/2023/2772/oj.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 120–127. ACM.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

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. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

A Held-out Results and Significance Tests

Table 2 gives the full held-out recall/MRR numbers summarised in §5; Table 3 gives the paired significance tests for the main benchmark and the held-out company.

method	R@5	R@10	MRR
BM25	0.406	0.594	0.269
LLM-QE	0.422	0.656	0.291
LLM-QE$_{\text{SBERT}}$	0.438	0.641	0.310
CE-rerank(BM25)	**0.547**	0.734	0.334
CE-rerank(Hyb-SBERT)	0.531	0.750	0.333
OAE-lite(LSA)	0.516	**0.812**	0.333
OAE-lite(SBERT)	0.500	0.766	**0.349**

Table 2: Held-out zero-shot retrieval on Company F (n=64), unseen in ontology construction or pooling design. Best per column in bold.

comparison	R@5	R@10	MRR
OAE-lite vs. Hyb-LSA	**.0009**	**.019**	**.014**
OAE-lite$_{\text{S}}$ vs. Hyb-SBERT	**.015**	**.0002**	**.042**
OAE-lite$_{\text{B}}$ vs. Hyb-BGE	**.019**	**.043**	**.0015**
OAE-lite vs. PRF-RM3	.071	**.0002**	.072
OAE-lite$_{\text{S}}$ vs. BM25	**.008**	**.0025**	.151
OAE-lite$_{\text{S}}$ vs. CE(BM25)	**.0003**	**.0023**	**.006**
OAE-lite$_{\text{S}}$ vs. CE(Hyb-SBERT)	**.0001**	**.0002**	**.009**
Controls (outside the Holm family)			
Hyb-LSA vs. BM25 (fusion only)	1.0	1.0	.65
LLM-QE vs. BM25	**.008**	**.029**	.062
LLM-QE vs. PRF-RM3	.071	**<.0001**	**.030**
OAE-lite vs. LLM-QE	1.0	1.0	.69
OAE-lite$_{\text{S}}$ vs. LLM-QE$_{\text{S}}$	1.0	.30	.75
Held-out Company F (n=64)			
OAE-lite(LSA) vs. BM25	.144	**.0094**	**.0499**
OAE-lite(S) vs. BM25	.238	**.0266**	**.0272**
LLM-QE vs. BM25	1.0	.455	.459
OAE-lite(LSA) vs. LLM-QE	.109	**.041**	.234
OAE-lite(S) vs. LLM-QE$_{\text{S}}$	.388	.057	.099
OAE-lite(S) vs. CE(BM25)	.607	.824	.749
OAE-lite(S) vs. CE(Hyb-SBERT)	.791	1.0	.741

Table 3: p-values: exact McNemar (R@5, R@10) and sign-permutation on reciprocal ranks (10^5 resamples, fixed seed; MRR). Bold: $p<.05$. In the first block, all five MRR comparisons except best-vs-BM25 (the three scorer-wise expansion tests and the two CE tests) survive Holm correction as a combined family; the control and held-out blocks are reported uncorrected and are not part of that family.

B Ontology Construction

We designed the ontology as a lightweight, auditable retrieval artifact rather than a full OWL/SHACL implementation. Each node stores requirement labels, expected evidence phrases,

fields, units, and validation notes used to generate query variants; the validation rules are SHACL-like and implemented in code.

Source materials. (i) The ESRS E1 requirements and their implementation guidance (European Financial Reporting Advisory Group (EFRAG) and European Commission, 2023), including the datapoint field names. (ii) IFRS S2 and SSBJ climate mappings, held as *reference-only* pointers for cross-standard transfer; they contribute terms only to OAE-FULL. (iii) General domain vocabulary (GHG Protocol scope terminology, common units). Company reports were *not* used to author the core ontology terms, and the held-out company was excluded entirely.

Workflow. (1) Select the 12 ESRS E1 evaluation requirements spanning transition plan (2), targets (4), energy (2), GHG emissions (3), and gross/removals separation (1). (2) Normalise each requirement into one node with: `label`, `topic`, `required_fields`, `optional_fields`, `allowed_units`, `allowed_values`, `synonyms`, `evidence_queries`, `standard_mappings`, and `validation_notes`. The `evidence_queries` are 2–3 short phrases naming how the evidence is typically *titled or tabulated* in a report (e.g. for gross Scope 2: "Scope 2 market-based location-based emissions"; "Scope 2 GHG emissions tCO2e"; "purchased electricity emissions Scope 2"). (3) Derive the three expansion tiers: LITE = label + `evidence_queries`; FIELD = lite + required/optional field names; FULL = field + synonyms + units/values + cross-standard mappings. Adding 13–24 terms, LITE sits inside the same budget as PRF-RM3's 20 feedback terms and LLM-QE's 20-term cap.

Guardrails against leakage. No gold label was consulted while writing expansion terms; terms are grounded in the standard and in general domain vocabulary, never in the content of a target page. The held-out company took no part in ontology authoring, hyperparameter choices, or candidate pooling, and its candidate pool is the union of every compared method's top-k. We report the residual risk that the main benchmark's pool vocabulary overlaps OAE terms more than BM25's (Jaccard 0.87 vs. 0.79) in Limitations.

Maintenance. Because every field is explicit and versioned, a standards change can be implemented as a localised edit and audited through version-controlled diffs. Governance of human-reviewed additions — who may add an evidence phrase, and on what basis — remains future work.

C LLM-QE: Prompt and Protocol

The expander is prompted once per requirement, with no corpus, gold labels, or ontology content in scope. Exact prompt:

```
  Standard: ESRS E1 (Climate Change).
Disclosure requirement: {requirement_text}
Task: produce search-query expansion terms that
would help a keyword-and-embedding retrieval
system find the page or section of a corporate
sustainability report that contains the evidence
for this requirement.  Output at most 20
whitespace-separated terms or short phrases.
Output only the terms, no explanation, no
punctuation other than what appears inside a term.
```

The retrieval query is `requirement_text` followed by the returned terms (truncated to 20), scored by the same HYB-LSA and HYB-SBERT scorers used for OAE-lite. The generated expansions were frozen before retrieval evaluation, so the reported runs use deterministic queries. Example output for gross Scope 2: `Scope-2 gross indirect emissions market-based location-based purchased-electricity purchased-heat steam cooling tCO2e dual-reporting grid-emission-factor residual-mix contractual-instruments` . . .

D Corpus Composition

company	subset	chunks	artifacts
Company A	EU/CSRD	451	PDF
Company B	US	128	PDF, HTML
Company C	US	262	PDF
Company D	US	212	PDF
Company E	EU/CSRD	124	PDF
Total (5 companies)		1,177	

Table 4: Real-artifact corpus. Company F, held out from ontology construction and tuning, adds 158 ESRS-native pages.

E Which Expansions Help, and Which Hurt

This section reports per-requirement Recall@10 for BM25 and OAE-lite on the 5-company benchmark and explains the two largest gains and the two weakest cases. The pattern is consistent: ontology phrases help where the requirement's regulatory wording and the report's own table/section wording genuinely diverge, and hurt where a requirement has *several* gold pages that the sharpened query collapses onto one.

Success: carbon credits / gross separation (n=5, R@10 .60 → 1.00). Helpful terms: gross emissions, carbon credits, offsets, removals, retired credits, net emissions. Retrieved evidence: credits table, methodology note. The requirement asks whether gross figures are kept *separate* from removals and credits. Companies title these pages "Carbon credits" or "Beyond value chain mitigation"; the raw requirement text matches the main GHG table instead. Ontology credit/removal vocabulary surfaces the credits page in its own right.

Success: target year (n=19, R@10 .47 → .84). Helpful terms: target year, milestone year, 2030, 2050, net zero by 2050. Retrieved evidence: target table, target narrative. "Extract the target year or milestone year" shares almost no surface form with report prose, which writes "by 2030" or "2050 net zero". Explicit milestone tokens bridge the mismatch — the largest single-requirement gain in the benchmark.

Failure: target statement (n=16, R@10 .94 → .75). Harmful terms: net zero target Scope 1 Scope 2 Scope 3, science-based target. Retrieved evidence: target narrative (wrong sibling page). The only regression. Where a company states targets across several pages (all of them gold), the added target vocabulary concentrates the ranking on one canonical target-summary page: for Company C the gold pages fall from rank 5 and 6 to 17 and 16 while an adjacent target page takes rank 1. Expansion sharpens precision at the cost of multi-gold coverage.

Weakest case: target base year (n=17, R@10 .53 → .65). Ambiguous terms: base year, baseline year (year tokens carry little discriminative signal). Retrieved evidence: ESG data table (wrong: reporting-period years). Improves but stays the weakest requirement. Year-heavy queries retrieve year-dense data tables whose columns are *reporting* periods rather than the page tying a base year to a target — the "reporting-period year vs. base/target year" type in the error taxonomy (§6). Year terms need target context to disambiguate.

The residual "transition plan vs. generic sustainability strategy" type also persists for the LSA scorer (TP existence .50 → .67) but is resolved by the SBERT scorer (→ 1.00), indicating a scorer rather than a vocabulary limitation.

The Climate Contracting Knowledge Graph: Construction and Evaluation

Georgia Ray[1,2], **Felix Cohen**[1]
[1]The Chancery Lane Project, [2]Imperial College London
Correspondence: georgia.ray@chancerylaneproject.org

Abstract

Contracts serve as a cornerstone mechanism of private environmental governance, and yet their form as unstructured text has meant a recent overhaul in their construction and iteration due to large language models (LLMs). As such, historical forms of knowledge sharing in the legal field must also change. We present the Climate Contracting Knowledge Graph, a publicly released resource that organises the model clauses, glossary terms, and guides of The Chancery Lane Project (TCLP) into a queryable property graph, linked to the laws and standards they invoke. Existing climate knowledge graphs target policy documents or climate-model metadata; none represent the contractual instruments through which private climate commitments are implemented. The graph is built with a lawyer-in-the-loop pipeline and achieves a statistically significant improvement in mean average precision over dense and lexical baselines. The graph, models, and a lightweight retrieval service are released publicly.

1 Introduction

Private environmental governance (PEG) is the private implementation of standards, monitoring, enforcement, and dispute resolution toward public ends (Vandenbergh, 2013). Vandenbergh & Johnson characterise contracting and other private enforcement mechanisms as "an important gap-filling effort given the limits of international, national and sub-national climate mitigation" (Vandenbergh and Johnson, 2021). Contracting, one component of PEG, has been shown to function as a lever in emissions reduction (Wang et al., 2024), and has been proposed as an enforcement mechanism for business-to-business climate action (Leonard, 2020). More recently AI companies have been called to provide environmental disclosures, and scholars have recognised contractual lock-in of energy supply for large-scale infrastructure projects powering AI (Luccioni et al., 2024, 2025).

In order to practice PEG, lawyers must be able to locate and deploy the relevant climate-aligned contracting content. However, as legal practice increasingly incorporates large language model (LLM)-based workflows, conventional mechanisms of information retrieval are being displaced, and the infrastructure for hosting professional content must adapt accordingly. Knowledge graphs (KGs) offer a mechanism for embedding structured legal and climate knowledge within LLM-based systems, and such graphs are increasingly being constructed across the climate domain: for policy documents (Dutia et al., 2025), climate model data (Ceriani et al., 2026), and mapping climate science (Carbon Brief, 2026). However, to this point, none address contractual instruments through which much climate action is implemented.

We present such a resource: a knowledge graph constructed from professionally drafted and maintained clauses, glossary terms, and guides for climate-aligned contracting, linked to the laws and standards they invoke. The graph is constructed via a lawyer-in-the-loop pipeline that favours task-specific models where feasible and LLMs otherwise. A retrieval ablation indicates employing the KG structure contributes a discriminative signal beyond dense and lexical retrieval alone, notably improving mean average precision (MAP). We further demonstrate the resource's utility through potential applications and release a lightweight interface for integration with third-party LLMs via a Model Context Protocol (MCP) server.[1] The underlying data and models are also made publicly available.[2,3]

[1]https://mcp.chancerylaneproject.org/
[2]https://huggingface.co/thechancerylaneproject/legal-sustainability-ner
[3]https://labs.chancerylaneproject.org/project/tclp-clause-library/

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 124–131
October 28, 2026 ©2026 Association for Computational Linguistics

2 Related Works

The peculiarities of a specific domain language in climate science and law, and the challenges those peculiarities pose to natural language processing (NLP) have long been recognised. Webersinke et al. (2021) show that general language models represent climate text poorly and release ClimateBERT, while Chalkidis et al. release LEGAL-BERT as the legalese equivalent (2020). Both of these BERT-based applications have enabled further exploration of climate and legal language in assessing corporate climate risk disclosures (Bingler et al., 2022), detecting and assessing net zero targets (Schimanski et al., 2023), and predicting legal judgements (Niklaus et al., 2021).

Meanwhile, KGs are acknowledged as a tool to ground LLMs in structured, verifiable knowledge, improving factual reliability, reducing hallucination, and promoting responsible use (Agrawal et al., 2023; Pan et al., 2024). LLMs are based on a transformer architecture, which encodes contextual meaning into the numerical (vector) representation of language (Vaswani et al., 2017). However, this approach carries a quadratic computational cost that constrains context length (Tay et al., 2023), and has been empirically shown to attend to the full context non-uniformly, typically focusing on the beginning and end (Liu et al., 2023). As such, uncommon relationships, like those specific to climate legal practice, or context provided only once, are liable to be under-weighted or missed. KGs address this shortcoming by supplying relationships explicitly and in a machine-retrievable manner with lower noise and higher density than raw text (Edge et al., 2024).

Surveys of KG construction describe a common pipeline of extraction, entity resolution, and refinement, and note that several of these stages still require human input (Zhong et al., 2024; Hofer et al., 2024). Ours follows the same pipeline, using lawyer annotations to fine-tune the extractor, LLMs for typing and entity matching, and expert review at the refinement stage.

Our evaluation adopts the information-retrieval perspective on knowledge graphs (Reinanda et al., 2020), meaning we evaluate the graph as a structure for finding and ranking relevant documents and entities, not as a system that performs multi-step logical inference over its edges.

3 Presentation of the knowledge graph

3.1 Source corpus

The corpus comprises 200 climate clauses, 76 glossary terms, and 11 guides authored and maintained by TCLP.[4] Of the clauses, 18 currently have a "maintained" status and therefore have the richest metadata taxonomy. All clauses have jurisdiction and practice area tags, and some additionally carry sector tags. The entire corpus of TCLP content was included in creating the knowledge graph.

3.2 Construction pipeline

Data collection and annotation. To gather hand-annotations from non-technical legal professionals, we built a multi-user annotation platform, seeded by an out-of-the-box spaCy model. spaCy's own default predictions across the whole corpus were the starting point for legal review. Entities were then approved, adjusted, or removed by legal annotators, who were also able to add entities that had not been tagged by the out-of-the-box system. Entity labels were selected from a pre-approved set agreed with our internal legal professionals. Annotators were trained on the annotating platform and entity tagging in two 45-minute training sessions, led by report authors.

New entities were most often added where the general-purpose model had no matching category or did not understand the relevance of a given entity in the climate law context. For example, 'Paris' was tagged by spaCy as a location. Our annotators revised it as to be part of the broader phrase 'Paris Agreement' and changed the label type to 'Standard'. Chunks were prioritised for annotation using a greedy minimum set-cover logic: the smallest chunk set covering the most not-yet-covered entities was repeatedly calculated and then presented first. This process continued until all pre-seeded entities had been seen.

Ultimately, 18 annotators annotated 831 of the 1,266 total chunks from the corpus. Annotators could merge or normalise entities manually, by

[4]These TCLP-specific terms are defined institutionally. Clauses are climate-aligned legal clauses one can use in commercial contracts and legal documents (e.g. Consolidated Sustainability Clauses for Supply Chain Contracts or Climate Change Due Diligence Questionnaire for Suppliers). Glossary terms are standard definitions to help one draft climate-aligned contracts (e.g. Alternative fuel or Residual emissions). And guides are documents that provide simple guidance for common scenarios and opportunities to use climate-aligned contracting (e.g. Introduce climate provisions to contract parties or Deliver a climate transition plan).



Figure 1: Static visualisation of sub-component of knowledge graph. This sub-component is focused around the glossary term "climate professional." The red dots are glossary terms, the purple dots are clauses, the brown dots are entities, light blue are guides, bright blue are laws or regulations, and darker blue are standards or frameworks. Individual node labels are not intended to be legible at this size. The figure is included to show the general form of the graph and the types of entity it contains. The visualisation is provided by the neo4j platform and viewer.

free typing the canonical form for an entity set but additional normalisation is performed by an LLM in graph construction (see Section 3.3).

Extraction and labelling. The results of the annotation process were used to further fine-tune the LEGAL-BERT baseline (Chalkidis et al., 2020), creating an extractor tailored to this task and expert judgement in this domain. The model was run against unseen test chunks, achieving an F1 of 0.69 with a precision of 0.70 and a recall of 0.68 on binary entity extraction. Augmenting this tailored model was a simple and strict keyword matcher for organisation-specific terms, which are crucial to TCLP but low-frequency in the training data. Traditional NER includes type labelling in addition to binary extraction. Our manually annotated data was labelled, but those labels were very sparse (there were only 46 instances of the 'sector' label and 195 of the 'law' label. That, as compared to over 3000 instances of both 'defined term' and 'concept' entailed strong class imbalance). Unsurprisingly, the NER model underperformed in learning these class imbalanced labels, particularly on the lowest-represented labels, and so the NER labels were not relied on. Instead, an LLM (Claude Haiku via OpenRouter) was prompted to label the extracted entities and was provided with domain-specific knowledge curated by the same expert lawyers regarding the proper usage of each label. Because the LLM is only tasked with labelling, not extraction, resource efficiency is improved and hallucinated entities are avoided. Ideally, using the LLM for labelling would have been unnecessary but as the labels are not an ontological organiser in the graph, a semi-automated approach demoting the labelling component to non-lawyer-reviewed was a prioritisation choice.

3.3 Graph creation and schema

The extraction pipeline writes its output to an intermediate SQLite storage, allowing the graph to be built and changed incrementally. When content was altered, a hash checkpoint was created to for the changes and extraction rerun for the altered content.

Each source document is represented by a clause, glossary term, or guide node. Alongside any taxonomy fields included in the specific piece of content, every node includes a dense embedding (BAAI/bge-base-en-v1.5, 768 dimensions).

Entities also become independent nodes. Every extracted entity is already keyed in SQLite to the document it was extracted from. To create entity nodes, we read this database row by row and either 1) create a new node if the entity is not yet in the graph or 2) merge that instance of an entity with the existing node. For merging, a normalised name match is tried first; where that fails, an LLM (Claude Haiku, via OpenRouter) performs a second, fuzzy pass that links acronyms, abbreviations,

	Count
Parameters	
CO_OCCURS_WITH threshold	$\geq$ 31 docs
Nodes	
Clause documents	200
Glossary terms	76
Guides	11
Laws & regulations	86
Standards & frameworks	202
Entity nodes (total)	5,424
of which: Concept	3,214
of which: Defined Term	1,563
of which: Object	1,128
of which: Org	223
of which: Standard	189
of which: Person	189
of which: Party to Contract	282
of which: Law	59
Edges	
CONTAINED_IN	16,366
CO_OCCURS_WITH	66,924
SIMILAR_TO	10,307
REFERENCES_LAW	183
REFERENCES_STANDARD	521

Table 1: Properties of the KG. Current counts for node types and edges, further disaggregated into entity subtypes (based on labels; a single entity may carry more than one label so the sub-totals sum is more than the entity total). The CO_OCCURS_WITH threshold of $\geq$ 31 shared source documents corresponds to the 75th percentile of the empirical co-document distribution. All counts reflect the live graph at time of writing.

and minor variants to their canonical form. Fuzzy-merge decisions were also reviewed iteratively during development and used to refine the matching rules, although not with a formal evaluation process. Every entity is linked to each of its sources of provenance via a CONTAINED_IN edge.

Other edges include REFERENCES_LAW, REFERENCES_STANDARD, CO_OCCURS_WITH, and SIMILAR_TO. Figure 1 visualises the content linked by these different edges.

The threshold, node counts, and edge counts for the current build are reported in Table 1.

LSRF process. Once the KG was created, we realised the metadata information about laws, standards, regulations, and frameworks (LSRF) was incomplete as it had been derived from the manually maintained metadata.

We proceeded with a two-pronged approach for completeness and verification. As a first step, we graduated entities labelled as 'law' or 'standard' from an Entity node type to a LawOrRegulation or StandardOrFramework node type, accordingly. Secondly, we used a two-step LLM procedure in which a first, smaller LLM (google/gemma-4-31b-it) was given a chunk of text and asked to do a first pass identifying whether any LSRFs were present. A more sophisticated LLM (Claude Haiku) was then asked to classify each as a law, a standard, or neither. Both outputs, from the entity graduation approach and the LLM approach, were then manually reviewed by a domain expert.

Using two methods provides a measure of cross-validation and improves coverage. Approval rates were comparable between the methods, especially when considering occurrence (77.8% LLM, 75.4% graduation). However, the LLM pipeline was stronger on distinct instruments (69.6% vs 58.9%) and recovered 70.9% of the graduation instruments while the reverse was only 31.0%. (see Figure 2). This pipeline, including the human review, is an effort at properly seeding the graph with many of the relevant LSRFs for this community. As the graph expands, the approval ratings and cross-validated coverage from the LLM offer the potential for it to be used automatedly without human review.

4 Retrieval evaluation

4.1 Ablation against taxonomy derived gold standard

An important part of validating the usefulness of a KG in the information retrieval tradition is to have some measure of retrieval, and demonstrate performance compared to other retrieval mechanisms. We assess different retrieval mechanisms against manually established connections between pieces of information. We consider two clause nodes relevant to one another if they share at least two tags pooled across the practice-area and sector facets.

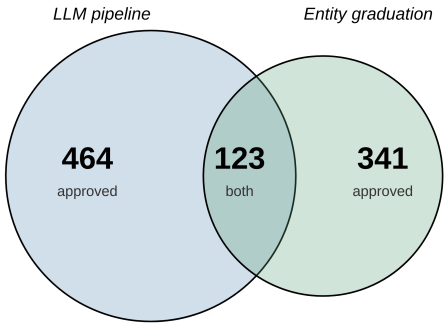
To measure retrieval, we report mean average precision (MAP) and mean reciprocal rank (MRR) as primary metrics and recall as secondary, with Recall@23 taken at the mean relevant-set size.

We compare four retrieval systems, fusing their components with reciprocal rank fusion (RRF, k=60) (Cormack et al., 2009): BM25 over clause text (Robertson and Zaragoza, 2009); dense retrieval over the clause embeddings; embeddings fused with BM25; and embeddings fused with the entity signal, in which two clauses are scored by the Jaccard overlap of the entities they share along CONTAINED_IN edges.

A random-ordering baseline has an expected precision of 0.12.

Augmenting dense retrieval with the entity layer

(a)



(b)

	LLM	Entity grad.
BY OCCURRENCE		
Suggested	596	452
Approved	464	341
Approval rate	77.8%	75.4%
BY DISTINCT INSTRUMENT		
Suggested	339	175
Approved	236	103
Approval rate	69.6%	58.9%
MUTUAL RECALL (DISTINCT INSTRUMENTS)		
Coverage of other pipeline	70.9%	31.0%

Figure 2: Coverage and overlap of LSRF identification pipelines. (a) Venn diagram showing approved occurrences (where occurrences are a node/clause pair rather than a unique node) for each pipeline. (b) Pipeline statistics by occurrence and by distinct instrument. Overlap was resolved using slug normalisation, parenthetical acronym extraction, and curated surface-form aliases, so that surface-form variants of the same instrument are counted as one match. Approval decisions reflect human review of pipeline-suggested instrument–clause links.

gives the strongest system on every metric, raising MAP from 0.465 to 0.513, a gain that is significant under a paired Wilcoxon signed-rank test ($p < 0.0001$). The entity-augmented system also significantly outperforms lexical fusion (embeddings + BM25; $p < 0.00001$), whereas BM25 fusion does not significantly improve on dense retrieval alone ($p = 0.062$). MRR is high and near-identical across the embedding-based systems ($\sim$0.81). This measure means a relevant clause is typically returned at rank one or two, no matter the system. The differences arise deeper in the ranking, which MAP and recall capture. The signal coming from the KG entities, therefore, is not reducible to lexical overlap. This finding indicates entities encode inter-document structure that neither dense nor lexical retrieval recovers. In short, using entities outperforms other methods of retrieval on all but one retrieval metric, suggesting there is an improvement in our ability to capture relevant information for these important private environmental governance questions using the KG.

4.2 Competency questions

On a more qualitative level, competency questions consider whether the graph can successfully answer the type of questions it is designed to answer (Grüninger and Fox, 1995).

We wrote a set of questions spanning the graph's intended uses, and confirmed that each is answerable by a single graph query. We had four types of questions: finding related clauses, concept and definition lookup, filtering by taxonomy, and lookup and aggregation. The full set and results can be seen in Appendix A.

5 Conclusions and Future work

We have introduced and publicly released the first, to our knowledge, knowledge graph for climate-aligned contracting, also demonstrating the application of a lawyer-in-the-loop pipeline for creating tools of this kind. Designed as a retrieval mechanism, retrieval ablation demonstrates that its entity layer significantly improves retrieval over dense and lexical baselines.

Potential applications of this graph include providing relevant information to practicing lawyers, as well as comparing instruments of private environmental governance to accepted standards, voluntary sustainability mechanisms, or public commitments.

A lightweight version of the retrieval interface is already released as an MCP server for integration with third-party LLMs, and as that tooling advances so too can this machine-readable climate-contracting knowledge graph advance with it.

Limitations

The principal limitation of this work is the availability of expert reviewers. We had hoped for three independent reviews at the annotation step but ultimately achieved only a single review of each chunk, so we cannot report inter-annotator agree-

ment. This scarcity of expert labels also left them too sparse to train a labelled entity extractor. As a workaround, we moved to binary entity extraction rather than labelled extraction and assigned types with an LLM. The fuzzy-merge decisions in Section 3.3 received the same single-reviewer assessment as the annotation step, and the results of that process are therefore also limited by the lack of inter-annotator review and full evaluation suite.

Similarly related to expert review, legal feedback on the results presented in Appendix A would have been an interesting extension of this project. As is, they serve only to demonstrate that these kinds of questions can be answered reasonably well, but we do not engage in a full conversation on how the legal community might receive them. This limitation is doubly challenging as experts would require not only expert legal knowledge, but also expert knowledge of the specific content used to create this KG. The same gap extends to the annotation process itself: we did not conduct qualitative interviews as to the annotators' experience, including the usefulness of pre-seeded entities or annotation guidance, so that qualitative account is missing although acknowledged as potentially rich.

Expert time was also a constraint on maintained content. Only 18 of the clauses had been actively maintained, and only these carry the associated metadata, so we could not leverage that metadata for categorisation and exploration across the whole graph. Relatedly, we treated clauses as connected to one another in assessing the retrieval performance of the KG based on the practice areas or sectors they share. This treatment was a by-product of how previous maintenance had been performed, but more time and more access to legal expertise would have allowed for a curated test set based on what legal practitioners think should be related.

Finally, data availability is a structural limitation of the legal domain. Because the methods of private environmental governance are typically proprietary, future applications of this work will be constrained by access to appropriate legal data.

Acknowledgements

We would like to thank the legal staff at The Chancery Lane Project who were involved in annotating, metadata tagging, and legal guidance including Humzah Khan, Clarissa Lambourne, Simone Potter, Natasha Morgan, Meena Kamath, Becky Annison and Charlotte Goosey. We also received annotation help from the excellent team at LexisNexis.

We also thank the TCLP staff and the many pro bono lawyers and other professionals who have drafted, reviewed and maintained the clauses, glossary terms and guides since the project began. That body of work is the corpus on which this knowledge graph is built.

We further thank Philip Detje who prototyped the application of this work for tenders, and was involved in conversations surrounding this draft.

The authors are affiliated with The Chancery Lane Project, the non-profit whose clauses, glossary terms, and guides constitute the source corpus and the resource evaluated in this paper. The authors declare no other competing interests.

This work was completed with the aid of LLMs, including in code generation, report editing, and figure design.

We are grateful to funding received from the Patrick J. McGovern foundation which made this work possible.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2023. Can knowledge graphs reduce hallucinations in LLMs?: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024)*, volume 1, pages 3947–3960.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Carbon Brief. 2026. Project cosmos. https://interactive.carbonbrief.org/cosmos/index.html. Accessed June 25, 2026.

Miguel Ceriani, Fiorela Ciroku, Alessandro Russo, Massimiliano Schembri, Fai Fung, Neha Mittal, Vito Trianni, and Andrea Giovanni Nuzzolese. 2026. The climate change knowledge graph: Supporting climate services. https://arxiv.org/pdf/2602.19786. Accessed June 25, 2026.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms

Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kaylan Dutia, Anne Sistema, Julie Saigusa, and Harrison Pim. 2025. Climate policy radar's open knowledge graph. In *NeurIPS 2025 Workshop on Tackling Climate Change with Machine Learning*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, et al. 2024. From local to global: A graph RAG approach to query-focused summarization. https://arxiv.org/pdf/2404.16130. Accessed June 25, 2026.

Michael Grüninger and Mark S. Fox. 1995. Methodology for the design and evaluation of ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI*, Montreal. Accessed July 14, 2026.

Marvin Hofer, Daniel Obraczka, Alieh Saeedi, Hanna Köpcke, and Erhard Rahm. 2024. Construction of knowledge graphs: Current state and challenges. *Information*, 15(8):509.

Louis Leonard. 2020. Under the radar: A coherent system of climate governance, driven by business. https://papers.ssrn.com/abstract=3598219. Accessed July 2, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment? In *2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024)*, pages 85–99.

Sasha Luccioni, Emma Strubell, and Kate Crawford. 2025. From efficiency gains to rebound effects: The problem of Jevons' paradox in AI's polarized environmental debate. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2025)*, volume 1, pages 76–88.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the 2021 Natural Legal Language Processing Workshop (NLLP 2021)*, pages 19–35.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Ridho Reinanda, Edgar Meij, and Maarten De Rijke. 2020. Knowledge graphs: An information retrieval perspective. *Foundations and Trends in Information Retrieval*, 14(4):1–159.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 15745–15756.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6).

Michael P. Vandenbergh. 2013. Private environmental governance. *Cornell Law Review*, 99(1). Accessed June 25, 2026.

Michael P. Vandenbergh and Bruce M. Johnson. 2021. The role of private environmental governance in climate adaptation. *Frontiers in Climate*, 3:715368.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. Accessed July 2, 2026.

Yangyang Wang, Lan Fang, and Mingchun Jiang. 2024. Optimal contract design for carbon emission reduction in a green supply chain: Monetary incentive vs. symbolic incentive. *Managerial and Decision Economics*, 45(5):3152–3162.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. ClimateBert: A pretrained language model for climate-related text. *SSRN Electronic Journal*.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2024. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62.

A Competency questions

Each question is one a practitioner might bring to the resource, paired with the graph operation that answers it and a representative result from the current build. The questions fall into four groups.

A.1 Finding related clauses

A1. Related clauses by shared entities *Question. Which clauses are most closely related to a given clause on the basis of shared domain entities alone?*

Answered by ranking clauses on the Jaccard overlap of the entities they share through CONTAINED_IN edges, with no embedding signal.

Querying **"The Net Zero Standard for Suppliers"** returns a coherent set of supply-chain and emissions clauses, led by "Target Product Carbon Footprint (Schedule for Consumer Goods Contracts)" and "Green Supplier Agreement Terms."

A2. Substantively related clauses *Question. Which clauses are most related in substance, combining meaning with shared entities?*

Answered by reciprocal rank fusion (k=60) of embedding cosine similarity and entity overlap. Querying **"Carbon Neutral Agreement Clause"** returns semantic neighbours that need not share exact terminology, including "Ambitious Carbon Neutral Clause" and several green shareholders' agreements.

A.2 Concept and definition lookup

A3. Concept lookup *Question. Which clauses mention a given concept?*

Answered by matching Entity nodes to clauses through CONTAINED_IN. **"Science Based Targets"** resolves to 31 clauses, "Scope 3 emissions" to 5, and "Net Zero" to 4, with distinct surface forms of a concept tracked separately.

A4. Defined term and glossary definition *Question. Where does a defined term appear, and what is its definition?*

Answered by traversing GlossaryTerm to ClauseName through CONTAINED_IN. **"Net Zero Target"** returns its glossary definition together with the clauses that use it, such as "Board Paper Implementing Net Zero for SMEs."

A.3 Filtering by taxonomy

A5. Taxonomy-scoped retrieval *Question. Which clauses match a combination of taxonomy filters?*

Answered by filtering on the slug-array properties of ClauseName. **Supply contracts whose climate solution is reducing greenhouse gas emissions** return two clauses: "Consolidated Sustainability Clauses for Supply Chain Contracts" and "Gain-Share Incentive for Reducing GHG Emissions." When queried for a combination absent from the corpus, the same operation returns an empty set rather than a spurious match, so the graph reports coverage gaps faithfully.

A.4 Reference lookup and aggregation

A6. Reference lookup *Question. Which clauses invoke a given law or standard?*

Answered by REFERENCES_LAW and REFERENCES_STANDARD edges. The **Paris Agreement** is invoked by 45 clauses, the **GHG Protocol** by 32, and **ISO 14001** by 3.

A7. Reference aggregation *Question. Which instruments are most referenced across the corpus, and how do curated and machine-extracted references compare?*

Answered by aggregating reference edges by target and by provenance. The most-referenced instruments are the Paris Agreement (45), the GHG Protocol (32), and the Science Based Targets initiative (25), followed by a long tail of single-clause references. The large majority of these edges originate from the human-reviewed LSRF classifier rather than the original curated metadata.

Beyond Dictionaries: Machine-Learned Normalization of Emission Units in Information Extraction

Lisa Reichenbach[1,*], Malte Schierholz[2,3], Anna Steinberg Schulten[2,3], Simon Oehler[1]

[1]Deutsche Bundesbank,
[2]Ludwig-Maximilians-Universität München (LMU Munich),
[3]Munich Center for Machine Learning (MCML)

Abstract

Sustainability reports denote greenhouse gas emissions with heterogeneous unit strings, such as abbreviations, inconsistent formatting, and scale markers. In emission extraction pipelines, these units must be normalized to a common target, e.g. `t CO2e`, to enable reliable comparison and aggregation of emissions across companies, reporting years, scopes, and regions. We propose a hybrid normalization approach: If a raw unit string is not found in a curated dictionary, a machine learning classifier maps noisy unit strings to normalized emission-unit categories that can then be converted to the target unit. The classifier is trained using character-level TF-IDF, lightweight string features, and LLM-assisted data augmentation. On random-splits of the augmented dataset, an ensemble model achieves approximately 95% accuracy. Additional evaluation on rare real-world unit strings indicates that supervised fallback classifiers can handle out-of-dictionary variants in climate-related information extraction.

1 Introduction

Corporate sustainability disclosures are key sources for assessing greenhouse-gas emissions, climate-related risks, and decarbonization progress. As companies disclose growing volumes of Scope 1, Scope 2, and Scope 3 emissions under frameworks such as the GHG Protocol, TCFD, and CDP (GHG Protocol, 2011; TCFD, 2017; CDP, 2024), traditional NLP- and LLM-based systems are increasingly used to extract structured climate-related information from these reports (Bingler et al., 2022; Auzepy et al., 2023; Dave et al., 2024).

A persistent challenge in downstream analyses such as benchmarking, time-series comparisons, and total-emission calculations is that extracted emission-unit strings vary in language, capitalization, punctuation, abbreviation, and scale notation.

Raw unit string	Norm. class	Factor
"Tonnes CO2e"	`t CO2e`	1
"mt CO2e"	`t CO2e`	1
"kg CO2 equivalent"	`kg CO2e`	0.001
"kilotonnes CO2e"	`kt CO2e`	1,000
"1,000 t CO2"	`kt CO2e`	1,000
"million t CO2e"	`Mt CO2e`	1,000,000
"US tons CO2e"	`USt CO2e`	0.907185
"kg CO2/kWh"	`Other`	–

Table 1: Examples of raw unit strings, normalized classes, and deterministic conversion factors to `t CO2e`.

As Table 1 illustrates, superficially similar strings may imply conversion factors that differ by several orders of magnitude. Extracted quantities must therefore be normalized to a common target unit, typically tonnes of CO2-equivalent (`t CO2e`).

Emission-unit normalization can be viewed as a post-processing step after value-unit pairs have been extracted by an upstream RAG-based pipeline (Dimmelmeier et al., 2024; Zou et al., 2025). Dictionary-based normalization handles known variants precisely, but is brittle to unseen spellings, multilingual expressions, abbreviations, and extraction artifacts, reflecting known limitations of rule-based and static-dictionary normalization approaches (Sproat et al., 2001).

In our extraction setting, this brittleness is substantial: in a prior run over 1,760 corporate reports, exact dictionary lookup left 23.8% of extracted value-unit pairs unresolved. Moreover, 18.5% of all pairs were dictionary misses that could be manually assigned to specific emission-unit classes other than `Other`. This indicates that many unresolved strings are potentially convertible emission units and motivates a supervised fallback component for out-of-dictionary variants.

Yet unit harmonization is rarely treated as a standalone task in ClimateNLP, where prior work has focused mainly on document classification, climate-risk identification, metric extraction, and named entity recognition (Webersinke et al., 2021; Bingler et al., 2022; Dave et al., 2024; Pan et al., 2025).

* Corresponding author: lisa.reichenbach@bundesbank.de

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 132–142
October 28, 2026 ©2026 Association for Computational Linguistics

We address this gap by formulating emission-unit normalization as a supervised text classification task within a hybrid dictionary-classifier pipeline. Dictionary lookup identifies known units, while a lightweight fallback classifier normalizes unknown unit strings. Using a manually curated dictionary augmented with LLM-generated variants (Ding et al., 2024; Dai et al., 2025), we evaluate the approach across complementary evaluation settings covering both augmented and rare real-world unit strings.

2 Related Work

Normalization in information extraction is commonly treated as a post-processing step that maps heterogeneous surface forms to canonical representations. Unit and quantity normalization often relies on curated inventories, rule-based matching, and deterministic conversion tables, which work well for known surface forms but are less robust to noisy, abbreviated, or domain-specific variants (Sproat et al., 2001). Early work on text and temporal normalization illustrates this pattern: CTEMP (Mingli et al., 2005), for example, extracts and normalizes Chinese temporal expressions using a grammar- and constraint-rule-based chart parser. Related value-unit extraction has been studied in biomedical text, including the extraction of laboratory test names, numerical results, units, and contextual attributes from free text (Kang and Kayaalp, 2013). More recent normalization and quantity-extraction systems combine lexical, statistical, or neural components with unit lookup and conversion rules (Lange et al., 2023; Ashmawy et al., 2023; Foppiano et al., 2019; Almasian et al., 2023). However, unit normalization is usually treated as auxiliary, with limited evaluation of fallback mechanisms for out-of-dictionary unit strings.

Climate-related information extraction has primarily addressed climate-document classification, risk identification, metric extraction, and named entity recognition in corporate disclosures (Webersinke et al., 2021; Bingler et al., 2022; Dave et al., 2024; Pan et al., 2025). Emission-unit harmonization is often handled implicitly through static dictionaries or downstream conversion rules, despite the fact that rare or noisy unit strings can lead to large aggregation errors. We address this gap by evaluating a hybrid dictionary-classifier approach for normalizing greenhouse-gas emission units in sustainability-report extraction.

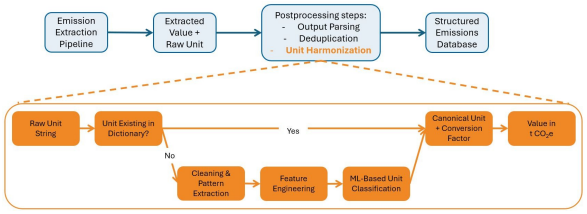


Figure 1: Overview of an emission extraction pipeline and the proposed hybrid extension for unit harmonization. In the evaluated extension, unresolved raw unit strings are passed to a machine learning fallback classifier before the extracted values are converted to t CO_2e and stored in a structured emissions database.

3 Methodology

3.1 Hybrid Harmonization Pipeline

Emission extraction pipelines commonly require a unit-harmonization step after raw value-unit pairs have been extracted. In a deterministic dictionary-based approach, extracted raw unit strings are matched exactly against a curated dictionary. If no match is found, the unit remains unresolved and cannot be reliably converted to a common target unit.

We propose and evaluate a hybrid extension in which unresolved raw unit strings are passed to a supervised machine-learning fallback classifier, as shown in Figure 1. The fallback is designed to recognize unseen lexical variants of known unit classes, while unit categories outside the predefined label space still require dictionary extension or model retraining. In this work, we evaluate the fallback mechanism offline as a standalone unit-normalization component.

3.2 Task Formulation

We formulate the fallback unit-harmonization component as a supervised classification task. Given an extracted raw unit string u that is not resolved by exact dictionary lookup, such as "million metric tonnes of CO2e", the model predicts a normalized unit class $c \in C$ such as `Mt CO2e` (see Table A.2 for complete class list). Each normalized emission-unit class, except non-convertible catch-all labels such as `Other`, is associated with a deterministic conversion factor f_c such that 1 unit of the normalized class = $f_c \times$ t CO_2e. For example, `Mt CO2e` corresponds to $f_c = 1{,}000{,}000$.

3.3 Data and Feature Representation

The dataset is based on a manually curated unit dictionary in which each instance corresponds to one distinct raw unit-string entry paired with its

normalized unit class and conversion factor. The dictionary also stores occurrence frequencies from previous applications of the emission-extraction pipeline to sustainability-report collections. These frequencies indicate how often a raw string was extracted across reports, reporting years, and scopes, but they do not create duplicate dataset instances and are used only for analysis, not as predictors. All frequency metadata were collected before the machine-learning experiments.

Given the highly imbalanced classes, as evident from the column "Manual" in Table A2, we augment underrepresented classes with GPT-4.1-generated variants (OpenAI, 2025), following prior LLM-based augmentation work under data scarcity and class imbalance (Ye et al., 2024; Ding et al., 2024; Dai et al., 2025). The augmentation is designed to expand existing semantic classes with plausible alternative surface forms, rather than to introduce new unit concepts. Because small changes in prefixes, denominators, or scale expressions can alter unit semantics, all generated candidates are deduplicated against the manual dictionary and manually validated. Lexical overlap among variants should be considered when interpreting performance estimates based on random data splits. The augmentation prompt is shown in Figure A1, and randomly selected examples are provided in Table A1 in Appendix A.1. The final dataset comprises 1,849 unit strings across eight target classes, including 471 manually validated LLM-generated variants. Table A2 reports the resulting class distribution.

The classifiers use only features derived from the raw unit string. Unit strings are lowercased, normalized, and represented using character-level TF-IDF features. Lowercasing was retained because it improved validation performance in preliminary experiments, despite discarding potentially informative case distinctions such as `mt` versus `Mt`; we revisit this trade-off in the error analysis and limitations. Additional regex-based indicators capture scale and quantity expressions such as *kilo*, *million*, *mega*, or *short tons*; simple numerical features such as string length are also included. Textual features are vectorized using character n-grams and combined with scaled numerical features in a `ColumnTransformer`. Target labels are label-encoded during training and decoded after prediction.

Full implementation details, including preprocessing steps, character n-gram settings, hyperparameter grids, selected model configurations, software versions, and random seeds, are provided in Appendix A.3. The code and unit dictionary are available in a public GitHub repository[1] to support reproducibility and future work.

3.4 Experimental Setup

As a trivial lower bound, we include a majority-class classifier that always predicts the most frequent normalized unit class in the training data, namely `Other`. Since exact dictionary lookup on the curated dataset is perfect by construction, we do not report it. We compare the majority-class classifier with several supervised models for the fallback task: Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, LightGBM, Multi-layer Perceptron (MLP), and a majority-voting ensemble classifier combining SVM, MLP, and Logistic Regression. Hyperparameters are optimized on the full dataset using grid search with stratified cross-validation.

Evaluation is conducted in two complementary settings. The **random-split dataset** evaluates model performance under a standard stratified split of the full dataset, including both manually curated and LLM-augmented unit strings. The **rare-unit dataset** is a robustness test split comprising manually curated unit strings whose exact raw surface forms occurred at most five times in the prior extraction runs, based on the frequency metadata defined in Section 3.3. The rare-unit test set was constructed from pre-existing dictionary entries and occurrence metadata, independently of model predictions and error analyses. LLM-generated variants are excluded because they were not observed in the extraction output and therefore do not have occurrence counts.

In the rare-unit experiment, the rare-unit test strings are excluded from model training, while LLM-augmented entries remain available in the training data to improve coverage of underrepresented classes. No additional hyperparameter tuning is performed. We instead use the hyperparameters selected in the preceding grid-search cross-validation experiments. Since that cross-validation used random splits of the full augmented dataset, some rare-unit strings may have contributed to hyperparameter selection. Thus, the rare-unit evaluation uses a fixed configuration and excludes the rare test strings from model fitting, but it is not

[1] https://github.com/gist-sustainability/unit_harmonization/releases/tag/v0.1.0

Model	10-fold CV			Random-split run		
	Acc.	$F1_w$	Macro F1	Acc.	$F1_w$	Macro F1
Majority-class	0.3067 ± 0.0026	0.1439 ± 0.0022	0.0587 ± 0.0004	0.3054	0.1429	0.0585
Logistic Regression	0.9524 ± 0.0026	0.9525 ± 0.0143	0.9621 ± 0.0103	0.9595	0.9596	0.9669
Random Forest	0.9421 ± 0.0189	0.9421 ± 0.0189	0.9551 ± 0.0134	0.9351	0.9355	0.9495
SVM	0.9516 ± 0.0116	0.9414 ± 0.0116	0.9542 ± 0.0068	0.9486	0.9489	0.9601
XGBoost	0.9448 ± 0.0149	0.9447 ± 0.0150	0.9570 ± 0.0107	0.9514	0.9515	0.9618
LightGBM	0.9486 ± 0.0165	0.9486 ± 0.0164	0.9598 ± 0.0144	0.9514	0.9515	0.9625
MLP	0.9492 ± 0.0103	0.9493 ± 0.0103	0.9603 ± 0.0081	**0.9649**	**0.9648**	**0.9751**
Voting Ensemble	**0.9551 ± 0.0114**	**0.9552 ± 0.0113**	**0.9671 ± 0.0082**	0.9595	0.9595	0.9708

Table 2: Offline evaluation of fallback classifiers for unit harmonization. Cross-validation results report mean ± standard deviation across 10 stratified folds on the full augmented dictionary-derived dataset. Random-split results are computed from a single stratified 80/20 train-test split on the random-split dataset.

fully independent of the preceding tuning procedure. Its purpose is to assess robustness to naturally observed low-frequency unit variants in the fallback use case.

We report accuracy, weighted F1-score ($F1_w$), and macro F1-score. In this single-label multiclass setting, accuracy is equivalent to micro-F1, so micro-F1 is not reported separately. Weighted F1 is used as the main imbalance-aware F1 metric because it accounts for class support and reflects the influence of frequent classes such as `Other`, `t CO2e`, and `kt CO2e`. Because aggregate metrics do not capture the practical severity of different error types, we complement them with per-class results and row-normalized confusion matrices, focusing on whether errors involve distant unit scales, such as `t CO2e` versus `Mt CO2e`, or conservative predictions of `Other`.

4 Results

4.1 Evaluation of Fallback Classifiers

Table 2 summarizes the offline performance of the fallback classifiers. We report 10-fold stratified cross-validation on the full augmented dictionary-derived dataset using `StratifiedKFold` with shuffling, fitting the complete preprocessing and classification pipeline separately within each fold. In addition, we evaluate each model in a single stratified 80/20 random-split run. Hyperparameters are fixed to the selected configurations reported in Table A4.

All supervised models clearly outperform the majority-class lower bound, whose weighted F1-score remains below 0.15. The Voting Ensemble achieves the strongest cross-validation performance across all metrics, with a weighted F1-score of 0.9552 ± 0.0113. Its comparatively low standard deviation also indicates stable performance across folds. On the random-split dataset, the ensemble remains among the best-performing models, with only MLP achieving higher performance in the single split run. Based on its consistently strongest cross-validation performance and competitive random-split results, we choose the Voting Ensemble as the selected fallback classifier.

The selected ensemble combines SVM, MLP, and Logistic Regression. It was chosen as a compact performance-runtime trade-off, since larger ensembles yielded only marginally higher validation scores while increasing runtime and complexity.

4.2 Per-Class Performance and Error Analysis

To interpret the aggregate results, we examine the per-class performance of the selected fallback classifier, the Voting Ensemble, on the test portion of the random-split dataset. Figure 2 shows the row-normalized confusion matrix for the emission unit categories with a strong diagonal, indicating high recall across classes.

Most remaining errors involve related unit categories or predictions of `Other`. This distinction is important because the practical severity of misclassifications is not uniform: predicting `Other` for a convertible unit mainly leads to a missed automatic conversion, whereas confusing distant scale classes, such as `t CO2e` and `Mt CO2e`, would directly induce errors of several orders of magnitude. The confusion matrix shows that such severe scale-related confusions are rare in the random-split dataset. The per-class scores in Appendix A.4 confirm this pattern, with all F1-scores above 0.92. This suggests that the fallback classifier is generally reliable for distinguishing emission unit categories.

The remaining errors in the random-split dataset

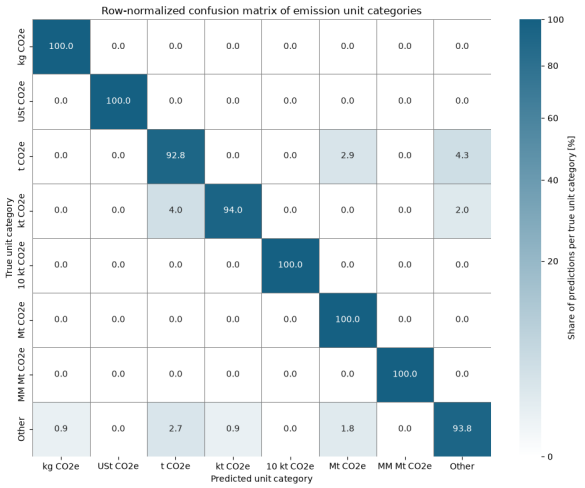


Figure 2: Confusion matrix of emission unit categories for the random-split dataset. Each row is normalized to 100%, showing the distribution of predicted unit categories for each true unit category. The diagonal entries correspond to the recall of the respective categories.

Raw string	True	Pred.	Error type
`mt CO2e`	`t CO2e`	`Mt CO2e`	case-sensitive scale ambiguity
`metric tons CO2 equivalent (based on 2016 data)`	`t CO2e`	`Other`	unit embedded in context
`emissions`	`Other`	`t CO2e`	no mass unit
`kgCO2e/sqm*year`	`Other`	`kg CO2e`	intensity instead of absolute unit
`th. t CO2 Eq.`	`kt CO2e`	`t CO2e`	abbreviated scale marker

Table 3: Representative misclassifications of the selected fallback classifier on the random-split dataset.

are concentrated in linguistically ambiguous or under-specified cases rather than in systematic confusions between distant unit scales. For example, capitalization-sensitive abbreviations can lead to errors: lowercase `mt` denotes metric tonnes and corresponds to `t CO2e`, whereas uppercase `Mt` denotes megatonnes and corresponds to `Mt CO2e`. As shown in Table 3, other errors involve abbreviated scale markers such as `th.`, units embedded in context, missing mass units, or boundary cases such as emission intensities.

4.3 Robustness on the Rare-Unit Dataset

As described in Section 3.4, the rare-unit test set consists of distinct manually curated real-world unit-string entries whose raw surface forms occurred at most five times in prior extraction runs. The resulting split is more imbalanced than the random-split dataset; the full train-test composition is reported in Appendix A.5, Table A6.

In this setting, the Voting Ensemble achieves an accuracy of 94.85%, a weighted F1-score of 94.84%, and a macro F1-score of 96.93%. The complete results are shown in Appendix Table A7. The weighted F1-score differs from the corresponding random-split result by only 1.11 percentage points, indicating strong performance on rare, naturally occurring unit variants rather than performance driven solely by frequent or LLM-augmented strings. As noted in Section 3.4, however, this evaluation is not fully independent of the preceding hyperparameter selection.

Detailed model comparisons, per-class scores, and the corresponding confusion matrix are provided in Appendix A.5. The rare-unit confusion matrix shows an even more conservative error pattern than the random-split evaluation: almost all remaining misclassifications are predictions of `Other`, while direct confusions between distant conversion-relevant unit scales are nearly absent.

5 Conclusion

This paper studies emission-unit normalization as a post-processing step for greenhouse gas emission extraction. We propose a hybrid approach that keeps deterministic dictionary lookup as a high-confidence first step and adds a supervised fallback classifier for out-of-dictionary unit strings. Using a curated and LLM-augmented dataset, we train different lightweight classifiers on character-level TF-IDF representations and simple domain-specific string features.

The results show that supervised classifiers can effectively extend dictionary coverage to unseen variants of known emission-unit classes, including rare real-world strings. The best-performing models achieve around 95% accuracy and weighted F1-score on both the random-split dataset and the rare-unit dataset. This suggests that the approach is not limited to frequent or augmented examples and can also handle rare real-world unit strings. Error analysis further suggests that most remaining mistakes stem from genuinely ambiguous abbreviations, missing unit information, or boundary cases such as emission intensities, rather than systematic large-scale conversion errors. The hybrid setup therefore provides a transparent and practical path toward more robust unit harmonization in climate-related information extraction.

6 Limitations

This study evaluates unit normalization as a standalone offline component rather than as part of an end-to-end emission extraction pipeline. Consequently, the evaluation does not capture interactions with upstream extraction errors, document context, table structure, or downstream aggregation. In practice, such errors may propagate, and scale confusions between units such as `t CO2e`, `kt CO2e`, and `Mt CO2e` can lead to conversion errors of several orders of magnitude.

A further limitation is the use of augmented dictionary-derived data. Although all LLM-generated variants were manually validated, random splits may overestimate generalization because highly similar surface forms can appear in both training and test data. The rare-unit dataset mitigates this by excluding LLM-augmented entries from the test set and evaluating only on rare manually curated real-world strings. However, it is not fully independent of model selection, as some rare-unit strings may have appeared in training or validation folds during the preceding random-split grid-search cross-validation. It may therefore be less conservative than a fully untouched test set. Moreover, the rare strings come from previous extraction runs and are not a prospective evaluation on newly extracted units from unseen reports. We do not include a mixed rare-plus-LLM holdout in order to keep the rare-unit test real-world-only, although such a synthetic holdout could be useful for sensitivity analysis on underrepresented classes.

Finally, the classifier is restricted to predefined unit classes and cannot handle genuinely new unit types without dictionary extension or retraining. Ambiguous abbreviations, multilingual expressions, and intensity-based units remain challenging. In particular, lowercasing improves robustness to inconsistent capitalization but removes informative distinctions such as `mt` versus `Mt`. Future work should integrate the classifier into the full extraction pipeline, incorporate document or table context, add case-sensitive or confidence-based rules for high-impact abbreviations, and use severity-aware metrics that weight errors by their implied conversion-factor difference.

Use of AI Assistants

AI-based writing assistants, including GPT-based tools, were used to support language refinement, wording suggestions, and minor editorial improvements during manuscript preparation. All AI-generated suggestions were manually reviewed, revised where necessary, and verified by the authors. The authors remain fully responsible for the content, analysis, results, and conclusions presented in this paper.

References

Satya Almasian, Vivian Kazakova, Philipp Göldner, and Michael Gertz. 2023. CQE: A Comprehensive Quantity Extractor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12845–12859, Singapore. Association for Computational Linguistics.

Mohamed Ashmawy, Mohamed Fakhr, and Fahima Maghraby. 2023. Lexical Normalization Using Generative Transformer Model (LN-GTM). *International Journal of Computational Intelligence Systems*, 16.

Alix Auzepy, Elena Tönjes, David Lenz, and Christoph Funk. 2023. Evaluating TCFD reporting—A new application of zero-shot analysis to climate-related financial disclosures. *PLOS ONE*, 18(11):1–23.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

CDP. 2024. CDP Corporate Questionnaire 2024. `https://www.cdp.net/en/disclose/question-bank`. Accessed: 2026-07-02.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Fang Zeng, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2025. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *IEEE Transactions on Big Data*, 11(3):907–918.

Aditya Dave, Mengchen Zhu, Dapeng Hu, and Sachin Tiwari. 2024. Climate AI for Corporate Decarbonization Metrics Extraction. Papers 2411.03402, arXiv.org.

Andreas Dimmelmeier, Hendrik Doll, Malte Schierholz, Emily Kormanyos, Maurice Fehr, Bolei Ma, Jacob Beck, Alexander Fraser, and Frauke Kreuter. 2024. Informing climate risk analysis using textual information - A research agenda. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 12–26, Bangkok, Thailand. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie

Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019. Automatic Identification and Normalisation of Physical Measurements in Scientific Literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, DocEng '19, New York, NY, USA. Association for Computing Machinery.

GHG Protocol. 2011. The Greenhouse Gas Protocol: A Corporate Accounting and Reporting Standard: Revised Edition. Technical report, World Resources Institute and World Business Council for Sustainable Development. Accessed: 2026-07-01.

Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of Pathology Informatics*, 4(1):23.

Lukas Lange, Jannik Strötgen, Heike Adel, and Dietrich Klakow. 2023. Multilingual Normalization of Temporal Expressions with Masked Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1174–1186, Dubrovnik, Croatia. Association for Computational Linguistics.

Wu Mingli, Li Wenjie, Lu Qin, and Li Baoli. 2005. Ctemp: A chinese temporal parser for extracting and normalizing temporal information. In *Natural Language Processing – IJCNLP 2005*, pages 694–706, Berlin, Heidelberg. Springer Berlin Heidelberg.

OpenAI. 2025. Introducing GPT-4.1 Model Family. Accessed: 2026-07-01.

Huitong Pan, Mustapha Adamu, Qi Zhang, Eduard Dragut, and Longin Jan Latecki. 2025. ClimateIE: A Dataset for Climate Science Information Extraction. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 76–98, Vienna, Austria. Association for Computational Linguistics.

Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of Non-Standard Words. *Computer Speech & Language*, 15:287–333.

TCFD. 2017. Recommendations of the Task Force on Climate-related Financial Disclosures. Technical report, Task Force on Climate-related Financial Disclosures. Accessed: 2026-07-02.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *CoRR*, abs/2110.12010.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. *ArXiv*, abs/2402.14568.

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.

A Appendix

A.1 Prompt Template and Examples for LLM-assisted Data Augmentation

To mitigate class imbalance in the unit normalization data, we use an LLM to generate additional variants for underrepresented unit categories. For each target category c, the model is prompted with examples from the well-represented `t CO2e` class and instructed to produce realistic unit strings with analogous surface-form variation patterns, including differences in spelling, formatting, abbreviations, whitespace, special characters, and capitalization. The prompt also restricts the output to variants without additional parenthetical descriptions and requests a Python list for straightforward downstream processing.

The prompt template used for the generation is shown in Figure A1. Here, c denotes the target unit category, while `{tco2e_variants_list}` denotes the example list from the well-represented `t CO2e` class.

Augmentation prompt

Please generate a list of 100 unique variants for the unit category c, following the style and diversity of the example list below for "t CO2e". The variants should differ in spelling, formatting, whitespaces, abbreviations, use of special characters and/or capitalization. Do not include any variants with additional descriptions in parentheses, e.g., "(per shipment)", "(biogenic)", etc.
Please avoid duplicates and ensure that all variants are realistic and could plausibly appear in real-world data. The output should be a Python list.
Here is my example list:
`{tco2e_variants_list}`

Figure A1: Prompt template used for LLM-assisted data augmentation. The placeholder c is replaced by the target unit category, while `{tco2e_variants_list}` is replaced by the example variants from the well-represented `t CO2e` class.

To illustrate the realized diversity of the augmentation, we manually inspected the LLM-generated variants from the final augmented dictionary. Table A1 presents randomly selected examples from different target classes and summarizes the main types of observed surface-form variation, including capitalization, hyphenation, abbreviations, lexical expansions, and Unicode characters. Although some variants remain close to existing dictionary entries, the examples demonstrate that the augmentation produces a range of plausible surface-form realizations within predefined semantic classes.

Target class	Generated variant	Observed variation
MM Mt CO2e	mm mt CO2-Equivalent	case, hyphen, expansion
kg CO2e	kg CO2-Equiv	hyphen, abbreviation
MM Mt CO2e	mm mt CO_2e	case, Unicode subscript
MM Mt CO2e	MMMt CO_2-equivalent	abbreviation, Unicode, hyphen
10 kt CO2e	10 KT CO2 equivalent	case, expansion
USt CO2e	us ton CO_2 eq	case, singular, abbreviation
kg CO2e	Kg CO2-Eq	case, hyphen, abbreviation
10 kt CO2e	10 kt CO2-Äq	hyphen, German abbreviation

Table A1: Randomly selected examples of manually validated LLM-generated unit-string variants. The final column summarizes the main types of observed surface-form variation.

A.2 Class Distribution by Data Source

Table A2 reports the class distribution of the final unit-string dataset across all eight target classes, distinguishing between manually curated entries and validated LLM-generated variants. The augmentation was applied selectively to classes with limited manual coverage, while already well-represented classes were kept unchanged. The category `Other` contains strings that cannot be unambiguously assigned to a target emission-unit class or refer to incompatible quantities, such as "emissions", "kWh", or "t CO2e/million euros".

A.3 Implementation and Hyperparameter Details

All models were implemented in Python version 3.11.9 using scikit-learn version 1.9.1, with XGBoost version 3.2.0 and LightGBM version 4.7.0. Unless stated otherwise, we used the default parameter settings of the respective libraries. Random seeds were fixed to 42 for data splitting and model training where applicable.

Unit class	Manual (Proportion)	Augmented	Total
t CO2e	343 (24.9%)	0	343
kt CO2e	249 (18.1%)	0	249
Mt CO2e	184 (13.4%)	0	184
kg CO2e	25 (1.8%)	101	126
USt CO2e	4 (0.3%)	112	116
10 kt CO2e	4 (0.3%)	116	120
MM Mt CO2e	2 (0.1%)	142	144
Other	567 (41.1%)	0	567
Total	**1378**	**471**	**1849**

Table A2: Class distribution of unit strings by data source. Manual entries originate from the curated unit dictionary, whereas augmented entries are validated LLM-generated variants.

Hyperparameters were selected on the training portion using grid search with 5-fold stratified cross-validation and weighted F1-score as the refit criterion. The selected pipeline was then refitted on the full training portion and evaluated on the held-out test set. Raw unit strings were represented by a concatenated text feature and one numerical length feature. The text feature combined: (i) a lowercased raw string in which numeric substrings were replaced by `NUMBER`; (ii) a regex-based quantity keyword covering markers such as *kilo*, *thousand*, *kt*, *million*, *mio.*, *mega*, *kg* and *US/short tons*; and (iii) a cleaned unit-base string obtained by removing numeric expressions and non-word characters, removing the token *of*, stripping whitespace, and lowercasing. We also tested case-preserving variants of components (i) and (iii), as well as variants preserving case in only one of the two components. These alternatives performed worse in preliminary validation experiments and were therefore not used in the final configuration.

The concatenated text field was vectorized using character-level TF-IDF with `analyzer=char_wb`, `ngram_range=(1,3)`, `min_df=3`, `max_df=0.95`, `max_features=1000`, and `sublinear_tf=True`; all other vectorizer parameters were left at their scikit-learn defaults, including L2 normalization. The numerical length feature was standardized using `StandardScaler`.

The complete hyperparameter search spaces are summarized in Table A3. For the Voting Ensemble, we first evaluated all base-model subsets of size three to six and then selected the final subset based on the trade-off between predictive performance, runtime, and model complexity. Table A4 reports the final model configurations used for the classifier comparison in the main experiments.

Model	Hyperparameter grid
Logistic Regression	Two grids were evaluated. First, C ∈ {0.01, 0.1, 1.0, 10.0, 100.0, 1000.0}, penalty ∈ {l2}, class_weight ∈ {None, balanced}, solver ∈ {lbfgs, newton-cg}, max_iter ∈ {1000, 2000, 3000}. Second, C ∈ {0.01, 0.1, 1.0, 10.0, 100.0}, penalty ∈ {l1, l2}, class_weight ∈ {None, balanced}, solver ∈ {saga}, max_iter ∈ {2000}
SVM	C ∈ {0.1, 1, 10, 100}, kernel ∈ {linear, rbf}, gamma ∈ {scale, auto}, class_weight ∈ {None, balanced}
Random Forest	n_estimators ∈ {100, 200, 500}, max_depth ∈ {None, 10, 20}, min_samples_split ∈ {2, 10}, min_samples_leaf ∈ {1, 4}, max_features ∈ {sqrt, log2}, bootstrap ∈ {True, False}, class_weight ∈ {None, balanced}
XGBoost	max_depth ∈ {4, 6}, learning_rate ∈ {0.2}, n_estimators ∈ {100, 300}, subsample ∈ {1}, colsample_bytree ∈ {0.8}, gamma ∈ {None, 0.3}, reg_alpha ∈ {None, 0.1}, reg_lambda ∈ {None, 1.5}
LightGBM	num_leaves ∈ {31, 63}, max_depth ∈ {−1, 6, 8}, n_estimators ∈ {100, 200, 600}, learning_rate ∈ {0.05, 0.1}, feature_fraction ∈ {0.8, 1.0}, bagging_freq ∈ {0, 1}, reg_lambda ∈ {0.0, 2.0}
MLP	hidden_layer_sizes ∈ {(100,), (200,), (100, 100), (200, 100)}, activation ∈ {relu, tanh}, alpha ∈ {$10^{-5}, 10^{-4}, 10^{-3}$}, learning_rate_init ∈ {0.001, 0.01}, solver ∈ {adam, lbfgs}, max_iter ∈ {500, 1000}, early_stopping ∈ {True, False}
Voting Ensemble	All base-model subsets of size 3 to 6 were evaluated from {XGBoost, LightGBM, Logistic Regression, Random Forest, SVM, MLP}, yielding 42 model combinations. The final subset, consisting of SVM, MLP, and Logistic Regression, was selected based on the trade-off between performance, runtime, and model complexity, as larger ensembles provided only marginal performance gains. For this subset, voting ∈ {hard, soft} was tested; weights=None.

Table A3: Hyperparameter grids used for grid search cross validation.

Model	Selected hyperparameters
Logistic Regression	C=100, penalty=l2, class_weight=balanced, solver=lbfgs, max_iter=1000
SVM	C=100, kernel=linear, gamma=scale, class_weight=None, probability=True, random_state=42
Random Forest	n_estimators=200, max_depth=None, bootstrap=False, class_weight=None, max_features=sqrt, min_samples_leaf=1, min_samples_split=2, random_state=42, n_jobs=-1
XGBoost	n_estimators=100, max_depth=6, learning_rate=0.2, subsample=1.0, colsample_bytree=0.8, gamma=None, reg_alpha=None, reg_lambda=1.5, random_state=42, eval_metric=mlogloss
LightGBM	n_estimators=600, max_depth=6, learning_rate=0.1, bagging_freq=0, feature_fraction=1.0, num_leaves=31, reg_lambda=0.0, random_state=42, objective=multiclass, metric=multi_logloss, verbosity=-1
MLP	hidden_layer_sizes=(200, 100), activation=tanh, solver=lbfgs, alpha=0.001, learning_rate_init=0.001, max_iter=500, random_state=42, early_stopping=True
Voting Ensemble	voting=soft, weights=None; base models: SVM, MLP, Logistic Regression

Table A4: Selected hyperparameters used for the classifier comparison.

A.4 Per-Class Classification Results

Table A5 reports the per-class precision, recall, and F1-scores of the selected fallback classifier on the test portion of the random-split dataset. The results complement the aggregate metrics and the confusion matrix in Section 4.2 by showing that performance is high across the eight unit categories. Most classes achieve F1-scores close to or above 0.95, with perfect scores for USt CO2e and MM Mt CO2e. Class t CO2e exhibits the lowest F1-score, driven by moderately lower precision and recall compared to the other classes. kt CO2e shows the lowest recall. The Other class remains robust, suggesting that incompatible or non-convertible units are usually not forced into a specific emission-unit category.

Class	Precision	Recall	F1	Support
t CO2e	0.9275	0.9275	0.9275	69
kt CO2e	0.9792	0.9400	0.9592	50
Mt CO2e	0.9024	1.0000	0.9487	37
kg CO2e	0.9615	1.0000	0.9804	25
USt CO2e	1.0000	1.0000	1.0000	23
10 kt CO2e	1.0000	1.0000	1.0000	24
MM Mt CO2e	1.0000	1.0000	1.0000	29
Other	0.9636	0.9381	0.9507	113

Table A5: Per-class performance of the selected classifier on the test portion of the random-split dataset. Support denotes the number of test instances per class.

A.5 Detailed Results on the Rare-Unit Dataset

The rare-unit evaluation uses a fixed train-test split designed to assess robustness on naturally occurring low-frequency unit strings. The test set consists only of distinct manually curated real-world unit-string entries whose corresponding raw surface forms occurred at most five times in prior extraction runs. These rare-test strings were excluded from training. The training set contains the remaining manually curated entries and all validated LLM-generated variants. LLM-generated variants are not included in the rare-unit test set because this evaluation is intended to focus on naturally observed rare strings. Table A6 summarizes the resulting train-test composition by class.

Class	Manual total	Rare test	Manual train	Aug. train	Total train
t CO2e	343	68	275	0	275
kt CO2e	249	49	200	0	200
Mt CO2e	184	36	148	0	148
kg CO2e	25	19	6	101	107
USt CO2e	4	4	0	112	112
10 kt CO2e	4	1	3	116	119
MM Mt CO2e	2	1	1	142	143
Other	567	113	454	0	454
Total	1378	291	1087	471	1558

Table A6: Train-test composition of the rare-unit evaluation setting. The rare test set contains only manually curated real-world unit strings whose exact raw surface forms occurred at most five times in prior extraction runs. LLM-generated variants are excluded from the test set and used only for training.

Table A7 reports the performance of all candidate fallback classifiers on this rare-unit dataset. The split composition in Table A6 shows that the rare-unit test set is substantially more imbalanced than the random-split dataset, with very small support for USt CO2e, 10 kt CO2e, and MM Mt CO2e. The aggregate metrics should therefore be interpreted together with the per-class results below. The Voting Ensemble achieves the strongest results on this dataset and shows the smallest absolute difference in weighted F1-score relative to the random-split dataset. This is particularly relevant because the MLP achieved the highest score on the random-split dataset, but exhibits a larger drop on the rare-unit dataset. The results therefore support selecting the Voting Ensemble as the more robust fallback classifier.

Table A8 provides the corresponding per-class results for the Voting Ensemble on the rare-unit dataset. Performance remains high across the main classes, with F1-scores above 0.91 for t CO2e, kt CO2e, Mt CO2e, kg CO2e, and Other. The class Other also remains reliable, with a recall of 0.9735, suggesting that rare non-target or incompatible unit strings are usually not forced into one of the emission-unit classes. For classes with very small rare-test support, especially USt CO2e, 10 kt CO2e, and MM Mt CO2e, the perfect scores should not be overinterpreted. Overall, the per-class results indicate that the classifier generalizes well to rare real-world unit variants and that the strong aggregate performance is not solely explained by frequent classes or by evaluation on LLM-augmented strings.

Model	Acc.	$F1_w$	Macro F1	$\Delta F1_w$
Log. Regression	0.9416	0.9413	0.9620	0.0183
Random Forest	0.9141	0.9125	0.8228	0.0230
SVM	0.9175	0.9174	0.8260	0.0315
XGBoost	0.8900	0.8892	0.7701	0.0623
LightGBM	0.9038	0.9035	0.8600	0.0480
MLP	0.9347	0.9345	0.9593	0.0303
Voting Ensemble	**0.9485**	**0.9484**	**0.9693**	**0.0111**

Table A7: Robustness evaluation of the fallback classifiers on the rare-unit dataset. $\Delta F1_w$ denotes the absolute difference in weighted F1-score between the rare-unit dataset and the test portion of the random-split dataset.

Class	Precision	Recall	F1	Support
t CO2e	0.9254	0.9118	0.9185	68
kt CO2e	0.9400	0.9592	0.9495	49
Mt CO2e	1.0000	0.9167	0.9565	36
kg CO2e	1.0000	0.9474	0.9730	19
USt CO2e	1.0000	1.0000	1.0000	4
10 kt CO2e	1.0000	1.0000	1.0000	1
MM Mt CO2e	1.0000	1.0000	1.0000	1
Other	0.9402	0.9735	0.9565	113

Table A8: Per-class performance of the selected fallback classifier on the rare-unit dataset. Support denotes the number of test instances per class.

Figure A2 shows the row-normalized confusion matrix for the rare-unit dataset, analogously to Figure 2. Compared to the random-split dataset, the error pattern is more conservative: a lot of the remaining misclassifications are predictions of Other, rather than confusions between alternative specific unit scales. This is relevant because errors are not equally severe in downstream emission aggregation. A false Other prediction prevents automatic conversion and may require manual review, whereas confusing units with very different conversion factors, such as kg CO2e and Mt CO2e, would directly distort normalized emissions by several orders of magnitude. For example, the errors for kg CO2e are entirely assigned to Other. Direct confusions between specific emission units are rare and occur only in small proportions. This suggests that the classifier tends to avoid assigning a potentially

wrong conversion-relevant unit category when uncertain, reducing the risk of severe scale-related conversion errors.

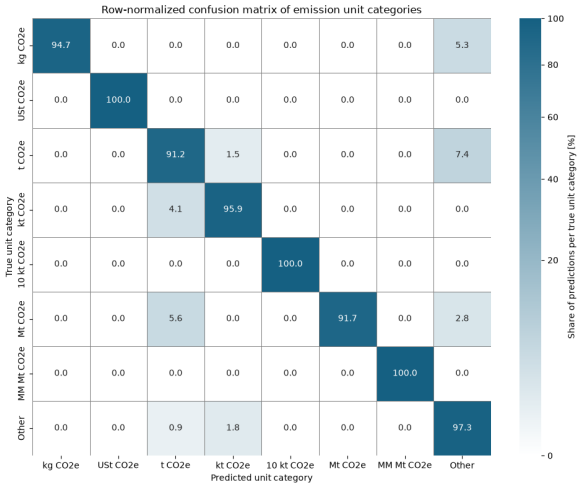


Figure A2: Row-normalized confusion matrix of predicted emission-unit categories on the rare-unit dataset. Values are percentages per true category; diagonal entries correspond to category-wise recall.

FoodClimNER: A Manually Annotated Corpus for Food Practice Entities Related to Climate Change

Mathieu Laï-king[1*], Louise Deléger[2], Arnaud Ferré[2], Yann Le Bodo[3], Estelle Chaix[1*]

[1]DER, ANSES, Maisons-Alfort, France
[2]MaIAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France
[3]DiSSES, ANSES, Maisons-Alfort, France
* Corresponding authors : {mathieu.laiking, estelle.chaix}@anses.fr

Abstract

Climate change is reshaping food systems and consumer practices, creating new challenges for food safety. Understanding how people modify their eating habits in response to climate disruptions and sustainability challenges is critical for assessing emerging microbiological risks. We present FoodClimNER[1], a new named entity recognition (NER) corpus containing extracts from open-access scientific articles on food science in the context of climate change, with manual annotations made by an expert. We define annotation guidelines with expert definitions of food practice entities and climate change entities. We evaluate baselines and a custom prompting method with recent LLMs on this task in a zero-shot setting. Results indicate that expert-level entity types and the presence of nested and discontinuous mentions make this a challenging testbed, with significant room for improving zero-shot methods. We also provide a detailed analysis of the importance of different parts of our custom prompt.

1 Introduction

Climate change is recognized as a major driver of changes in food safety challenges through its multiple impacts on food systems (Intergovernmental Panel on Climate Change (IPCC), 2023). It affects every stage of the food chain, from production and transportation to storage and food availability, while also modifying consumer food practices (Vermeulen et al., 2012). These interconnected changes are expected to increase the burden of foodborne diseases while reshaping food systems and consumption patterns worldwide (Dietrich et al., 2023). However, our understanding of how climate change affects consumer eating habits, the final link in the food chain, remains limited. Consumers are adapting their eating habits by adopting more sustainable dietary practices (e.g., reduced meat consumption, increased consumption of local, organic or plant-based foods) (Innocent et al., 2024) or in response to climate-related disruptions such as reduced availability of certain products or supply chain disturbances (Hadley et al., 2023). These shifts may introduce or amplify microbiological hazards, as new food practices emerge and foods with limited safety data enter consumers' diets (Duchenne-Moutien and Neetoo, 2021).

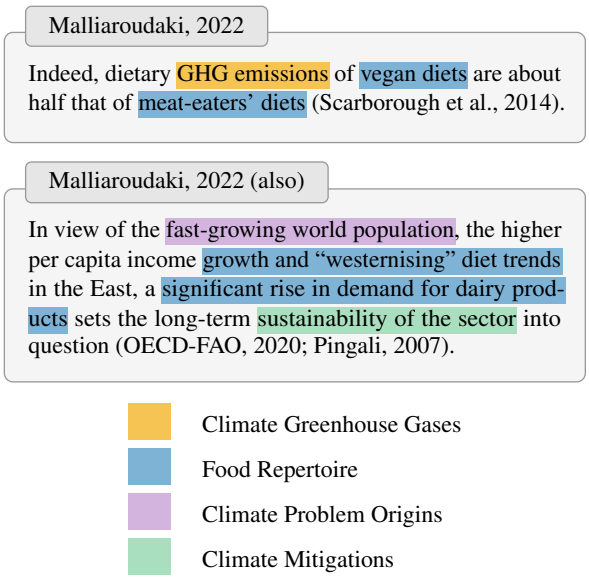


Figure 1: Examples of annotated sentences from FoodClimNER

Consumer food practices are documented across diverse text sources (scientific literature, grey literature, press articles, and social media), making natural language processing methods a promising choice for extracting structured information from them. Being able to extract information on food practices linked to climate change could help discover and understand emerging food practices that could pose a risk in the upcoming years due to

[1]We release our corpus publicly : https://huggingface.co/datasets/inrae-bibliome/foodclimner, as well as the code for our experiments : https://github.com/mathieulaiking/foodclimner

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 143–157
October 28, 2026 ©2026 Association for Computational Linguistics

climate change. For example, the transition from single-use plastic bags to reusable shopping bags and food containers, promoted to reduce environmental impacts, may unintentionally introduce new microbiological hazards. If these reusable items are inadequately cleaned or sanitized, they can act as vehicles for cross-contamination, particularly when used for high-risk foods such as meat, cheese or ready-to-eat products (Kasza et al., 2022).

However, annotating sufficient training data for such tasks remains both time-consuming and costly. The rapid advances in large language models (LLMs) over recent years (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Guo et al., 2025) offer a compelling alternative: these models have demonstrated strong zero-shot capabilities through a variety of prompting strategies (Kojima et al., 2022; Wei et al., 2023), thus reducing the need for large annotated corpora. These zero-shot capabilities also transfer to information extraction tasks such as named entity recognition (NER) (Sainz et al., 2024; Xie et al., 2023).

Here we define the term zero-shot as found in the literature, meaning the model is not trained (weight optimization) on the specific task directly. The model instead generalizes to the new task using the knowledge acquired during its training (pre or post-training) or its context (in the case of LLMs). For example, Sainz et al. (2024) include additional task information in their prompt (including entity types descriptions, examples, or annotation guidelines) and still consider this zero-shot. This use of zero-shot can be debatable (especially when using entity examples), but we do not show exact examples

Therefore, we present a new *zero-shot* evaluation corpus for food practices related to climate change. Our contributions include **FoodClimNER**, an expert-annotated NER corpus built from 10 full-text open-access scientific articles, with food practices and climate change entities (see examples in Figure 1), along with expert-defined annotation guidelines, and zero-shot experiments (2 baselines and a custom prompting method with recent LLMs) on this corpus.

2 Related Work

2.1 NER in the climate and food domain

There already exist several works on information extraction in the food domain. However, they are usually recipe datasets annotated with food entities and attributes such as quantities, units, and cooking processes (Popovski et al., 2019; Wróblewska et al., 2022) and do not include food practices or target other types of documents. In the climate domain, Bhattacharjee et al. (2024) introduce Climate Change NER as part of a broader suite of domain-specialized scientific language models, while Pan et al. (2025) propose ClimateIE, a corpus of climate science publications annotated for named entities, relations, and entity linking through an LLM-assisted annotation pipeline combined with expert validation.

Beyond NER, recent work has also explored domain-specialized LLMs for food and nutrition tasks, such as FoodyLLM (Gjorgjevikj et al., 2026), fine-tuned for nutrient estimation and ontology-based entity linking, alongside broader surveys on AI applications in food safety (Lopez et al., 2026).

As the majority of the existing NLP research in the food domain focuses on recipes, the originality of our approach is to annotate research articles in the food safety domain with food practices entities and link this to climate change entities.

2.2 Zero-shot NER methods

Zero-shot named entity recognition has attracted growing attention since the emergence of large language models. A first family of approaches, adapted across different model architectures, shares a common underlying idea: unifying multiple information extraction tasks within a single model by defining a unified output schema and training the model to generate it. This strategy was first explored with encoder-decoder architectures (Lu et al., 2022; Lou et al., 2023), later extended to decoder-only models (Wang et al., 2023), and more recently to encoder-only models (Zaratiana et al., 2025).

A second line of work distills LLMs into smaller, fine-tuned models by using the LLMs as automatic annotators over large text corpora. These approaches rely either solely on entity type labels (Zaratiana et al., 2023; Zhou et al., 2024) or additionally incorporate type descriptions (Aly et al., 2021; Cocchieri et al., 2025), computing similarity scores between sentence-token representations and entity-type (or type + description) representations.

Finally, a third category of approaches directly exploits the emerging capabilities of large language models, using original ways of defining annotation guidelines such as code (Sainz et al., 2024), tool-calling and multi-turn decomposition strategies (Xie et al., 2023), or retrieval-augmented cor-

rection of missing and erroneous entities (Cai et al., 2025).

3 FoodClimNER: A new corpus for food practices related to climate change

FoodClimNER is a new named entity recognition corpus on the subject of consumer food practices linked to climate change. The goal is to provide a testbed for detecting emerging food practices in scientific article documents that could present risks in the future due to climate change. However, this type of corpus is expensive to annotate, as we need expert annotators. Therefore, we aim to create a corpus used for the evaluation of modern zero-shot methods, as we believe our expert-defined entities and annotations will be a challenging task in this setting. Relation extraction was not performed; however, the co-occurrence of food practice and climate-related entities within the same textual context provides valuable information for identifying potential links between emerging food practices and climate change.

3.1 Annotation process

We describe our annotation process from schema creation to document annotation :

1. Two experts (in sociology and food microbiology) first developed the annotation schema for food practices. The annotation framework was designed through an iterative inductive-deductive approach, combining exploration of emerging food practices from the literature with theoretical frameworks describing food behaviors. Food practices were structured according to consumer food pathway, adapted from Desjeux's consumer itinerary framework (Desjeux, 1998). This structure was chosen to capture food practices at different stages of the consumer pathway and to support subsequent microbiological risk assessment. Climate-related entity types were selected from the schema of an existing climate change-related NER corpus (Bhattacharjee et al., 2024) and adapted to the objectives of this study. We focused on entities describing climate drivers, impacts and mitigation-related concepts so we filtered some relevant entity types from the corpus schema. We provide our annotation schema with detailed entity type definitions for food practices and climate change entity types in section 3.2.

2. Relevant document types included scientific articles, grey literature, press articles and social media content. Given the diversity of available sources, we focused the initial manual annotation corpus on open-access scientific articles to ensure reproducibility and sufficient contextual information. An exploratory literature search was conducted in PubMed and Scopus using a PECO-based framework (EFSA, 2010), combining terms related to climate change, sustainability, food, diets and consumer behaviors. The article selection was further informed by a broader food practice database built independently by three annotators from full-text readings of the screened corpus (Le Bodo et al., 2026). Articles containing at least one instance of a "food preparation" practice were prioritized, as this category was expected to offer a greater variability in surface shape ; two additional articles addressing food practices outside the French context were added among remaining eligible candidates to increase geographic diversity. The resulting ten articles span ten different journals across food science, nutrition, public health, environmental education, and sustainability disciplines.

3. Then, the full-text articles were manually reviewed by one expert, and some sections were discarded because they were judged to contain insufficient entity density or redundant information and also to alleviate the workload of the annotator, avoiding annotating the whole document. For each article the removed part length is very variable, for some documents almost nothing has been removed, and for some other we kept only the abstract or a small part of the complete paper.

4. Annotation was then performed by a single expert annotator, with expertise in food microbiology (PhD), climate change and text annotation (previous work annotating microbes entities in scientific articles). The software used for the annotation is the brat rapid annotation tool (Stenetorp et al., 2012).

The corpus captures two challenging linguistic phenomena: **nested entities**, where spans of different types overlap, and **discontinuous entities**, where a single logical entity spans non-contiguous

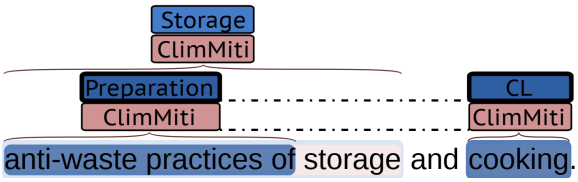


Figure 2: Example of nested + discontinuous annotations. We have one discontinuous mention of the "Food Preparation" entity: ["anti-waste practices of","cooking"]; that overlaps with one regular mention of the "Food Storage" entity: "anti-waste practices of storage". They are also both annotated as climate mitigations, as these practices serve to mitigate climate change.

text segments. We provide an annotated example in Figure 2.

3.2 Annotation Schema

We define two different main concepts of entities which are consumer food practice entities and climate change entities. Both of these concepts contain multiple subtypes of entities that have been manually annotated. We also consider a version where all the subtypes are directly considered as the main concept because we believe that extracting the information of having food practices as well as climate change entities in a given text is already valuable. This simplified version will be noted **FoodClimNER-simple** (results in appendix B).

The first macro-type of entities are consumer food practices (actions directly carried out by an individual or a household as a food consumer). The subtypes of entities are defined according to the food itinerary of the consumer, from planning to waste management:

1. **Food repertoire**: Covers practices and factors related to food choices (in terms of quality or quantity) that may or may not be part of an individual's diet (e.g., meat consumption, food choices).

2. **Logistics and Travel**: Includes pre-purchase practices (checking existing resources, meal planning, etc.) and travel to the place of purchase (using low-carbon transportation to go grocery shopping)

3. **Purchasing**: Covers purchasing practices (product selection, choice of retail locations, interactions with marketing) as well as non-market procurement strategies (self-production).

4. **Storage**: Covers food storage practices after purchase and before preparation (choice of packaging, adherence to or disregard of expiration dates, preservation, etc.)

5. **Food preparation**: All attitudes and practices related to food preparation and cooking (e.g., raw food diets, water conservation, reducing energy consumption, minimizing waste during preparation, etc.). It also includes skills and actions that we perform when cooking.

6. **Consumption** (table manners): Covers consumption practices related to eating and meals (e.g., portion sizes, choice of utensils).

7. **Waste Management**: Covers practices that take place after a meal, with consideration for the end of a food item's life cycle (storing leftovers, composting, sorting food waste, expiration date reasoning...).

Climate change entities encompass the phenomena, factors, impacts, causes, and strategies associated with climate change, as described in the text or recognized as falling within the realm of climate. They include environmental events, greenhouse gases, climate impacts, anthropogenic causes of climate change, and actions aimed at mitigating or responding to its effects. Climate change entities may be identified using domain knowledge. However, any causal or relational link between these entities and food practices must be explicitly supported by the text. The subtypes of entities are defined according to entities selected by the expert annotator from the Climate-change-NER corpus (Bhattacharjee et al., 2024), as well as an additional environmental concern type that was judged relevant by our expert:

- **Climate greenhouse gases**: Gases that cause heating of the atmosphere (e.g., "carbon dioxide", "CH4", "GHG emission")

- **Climate hazards**: Dangers with potential negative impact on climate, such as floods, wildfires, droughts, and heatwaves.

- **Climate impacts**: Effects of hazards, primarily negative effects on humans. We also consider climate impacts on livestock as impacts, as it indirectly affects humans.

- **Climate problem origins**: Problems that describe why the climate is changing. Key examples are fossil fuel and deforestation. We also

mention sectors that can be cited as causes of energy use. For instance, in a text about the energy consumption by the transport sector, the transport sector is annotated as a problem.

- **Climate mitigations**: Activities to reduce climate change or to better deal with the consequences.
- **Environmental concern**: This entity encompasses several aspects, e.g., pollution, biodiversity, waste, water, climate, etc.; it does not specifically refer to climate change; it is an intermediate/residual category within the environmental sphere for which the link to climate is not established or direct.

We also provide additional annotation rules in our annotation guidelines (Appendix A).

4 Experimental Setup

We want to evaluate two existing baseline methods and also a custom prompting method using recent state-of-the-art LLMs (baseline-comparable open-weight models and a large proprietary model) because the progress of LLMs is moving fast and the baselines may no longer be competitive even with task-specific fine-tuning. We also perform a pre-processing of the corpus before running the methods.

4.1 Corpus pre-processing

We ran the experiments on a pre-processed version of the corpus. The pre-processing is described here:

1. We replaced the references with a `[REF]` tag for single references and a `[MREF]` tag for multi-references, using regular expressions.
2. We split the documents into sentences using the spaCy sentencizer[2].

Even if the annotations are performed at the document level, we choose to split them into sentences, as some documents were long, and preliminary experiments showed that having a context too long was hurting performances for LLMs (possibly due to the lost-in-the-middle problem when giving large contexts to LLMs (Liu et al., 2023)). The citation replacement is principally added for easier sentence splitting and because we noticed a lot of multi-citations that did not add any useful context and were taking a lot of space in some sequences.

[2] https://spacy.io/api/sentencizer/

4.2 Existing Baseline Models

We select the following existing methods to have a lightweight encoder baseline and a fine-tuned LLM baseline:

- GLiNER (Zaratiana et al., 2023): It is a suite of encoder-based models where they fine-tuned a DeBERTa model using special `[ENT]` tokens that precede entity type labels on a large corpus with more than 13k different entity types automatically annotated using ChatGPT (`gpt-3.5-turbo-0301`). Then the special tokens and entities are concatenated with the sentence in which we want to find entities (using a special `[SEP]` token), and a dot product is computed between the input sentence token embeddings and the different `[ENT]` tokens for each entity type. We specifically use version 2.5 and the large variant (459M parameters) of the model (`gliner_large-v2.5`[3]).
- GoLLIE (Sainz et al., 2024): GoLLIE is a suite of decoder-based models fine-tuned on multiple information extraction tasks. The approach represents annotation guidelines as code syntax, allowing the model to use in-context learning to identify entity boundaries. The authors fine-tuned different sizes of CodeLLaMA models, transforming annotation guidelines (entity type labels, descriptions, and examples) in Python syntax, defining each entity type as a class. The input text is then appended as a Python string variable in the prompt, and the model generates the entity mentions as a structured list format. We use the 7B parameter (`GoLLIE-7B`) and 34B parameter (`GoLLIE-34B`) variants of the model (smaller and bigger one).

4.3 Custom Prompting Method

Because GoLLIE is based on an already outdated model and LLM capabilities have progressed significantly since then, we also use a custom prompting method so that we can evaluate more recent LLMs. We design a prompt that includes all the guideline information. We use one prompt for each document in the corpus, and we handle all entity types at once. Preliminary experiments strongly suggest that it helps disambiguate closely-related entity types (compared to handling one entity type

[3] https://huggingface.co/gliner-community/gliner_large-v2.5/tree/main

at a time), and it is also faster and cheaper as it divides the number of inferences by the number of entity types. Recent LLMs follow a system/user roles schema, so we design our prompt following this. Here is the composition of this prompt:

- For the system part: contains a basic persona (where we specify the model as expert in NER annotation)
- For the user part:
  1. Task instruction: simple sentence describing the NER task
  2. Entity types, names, and descriptions, and if available, entity mention examples
  3. Rules: Corpus-specific annotation guidelines if available, presented as a markdown section
  4. Output format: We require the output to follow a specific format, with an illustrative example provided. Here we only tried the JSON format.
  5. Then finally the input text from which we want to extract entities.

The full template is provided in Appendix D.

4.4 Models and configuration

We choose recent LLMs to evaluate this prompting method. We first choose two open-weight models that have a size comparable to the baseline LLM model (GoLLIE). We also add a large model that has state-of-the-art performance on text generation benchmarks, as well as a proprietary model:

- Gemma 4 (Team et al., 2026): Gemma 4 is a suite of vision-language models. We select the dense 31B parameter, instruction-tuned (`gemma-4-31B-it`) version of this model for our experiments.
- Qwen 3.5 (Qwen Team, 2026): Qwen 3.5 is also a suite of vision-language models. We try two variants of this model: the dense 27B parameter version (`Qwen3.5-27B`) and the dense 9B parameter version (`Qwen3.5-9B`). We choose these versions so that their size is close to the models used by the GoLLIE baseline.
- GPT 5.6 Sol (Singh et al., 2026): this is the latest[4] and (supposedly) largest variant of the GPT 5.6 model. According to its authors, it achieves state-of-the-art performance on many LLM textual benchmarks[5].

We run the models in greedy decoding, without thinking tokens, on the corpus sentences (after preprocessing the documents) so that the setting stays comparable with the GoLLIE baseline. It also gives a good idea of the performance of the recent models without having to use too much compute by using a multi-pass sampling evaluation and test-time compute.

We ran the open-weight models on single NVIDIA A100 SXM4 80 GB GPUs (except for GLiNER, which we ran on a local GPU NVIDIA RTX A2000 Laptop GPU). The proprietary GPT-5.6 model was ran using the OpenRouter API.

4.5 Evaluation

We perform the evaluation at the entity level and because the corpus has challenging mentions (discontinuous, nested entities), we use two modes of evaluation:

- **Partial**: we count a correct prediction if the predicted mention boundaries at least have an overlap with the true mention and if the predicted type is exact. This removes the discontinuous challenge; if there are multiple spans and the prediction overlaps with one of the spans, it will be counted as correct.
- **Strict**: we count a correct prediction if the boundaries strictly match (multi-spans for discontinuous entities must exactly match, so our methods do not handle this; all discontinuous mentions will be counted as incorrect) and if the predicted type is correct.

We then compute the micro-average for the metrics of precision, recall, and F1-score in both modes.

5 Results

We provide in this section our results on the FoodClimNER corpus; additional results on the simple variant of this corpus are included in appendix B.

5.1 Resulting corpus

We provide the statistics of our resulting corpus in Tables 1 (for general corpus statistics) and 2 (for the entity type number).

[4] at the time of our last experiments in July 2026

[5] `https://openai.com/index/gpt-5-6/`

Statistics	Value
Number of initial documents	10
Total nb. of words	13599
Number of sentences	527
Sent. avg length (char)	172.7 ± 88
Sent. avg length (word)	25.8 ± 13
Number of entity types	12
Number of mentions	1109
Number of unique mentions	727
Mention avg length (char)	26.7 ± 17
Mention avg length (word)	3.8 ± 3
Avg mentions per sentence	2.10
Discontinuous mentions	92 (8%)
Nested mentions	478 (43%)
Flat mentions	539 (49%)

Table 1: Corpus statistics for FoodClimNER

Entity Type	Count
Food Repertoire	297 (27%)
Food Preparation	132 (12%)
Food Supply Purchases	101 (9%)
Food Waste Management	72 (6%)
Food Storage	40 (4%)
Food Logistics Travel	36 (3%)
Food Consumption Uses	11 (1%)
Climate Mitigations	222 (20%)
Climate Problem Origins	136 (12%)
Environmental Concern	37 (3%)
Climate Impacts	19 (2%)
Climate Greenhouse Gases	6 (1%)
Climate Hazards	0 (0%)

Table 2: Entity type statistics for FoodClimNER

These statistics include sentence-level statistics, as our model inputs were sentences (split from the initial documents). We observe that annotated mentions are long (almost 4 words per mention in average), and that the corpus have a substantial part of nested entities (43 %, here a nested mention means that it has at least an overlap of one character with another mention). Our annotator did not identify any climate hazard entities in the selected documents, suggesting that the annotated texts primarily focus on other climate-related entities.

5.2 Overall results for both versions of our corpus

We provide the overall results (micro-average metrics of the different methods used) on the FoodClimNER corpus in Table 3.

First, looking at the baseline results, as expected, as we have expert definitions in our corpus, it is a really difficult task for GLiNER, which only supports entity type labels. GoLLIE performs better as it allows adding textual descriptions of each entity type, as well as examples. Similarly, our custom prompt containing information about the task (entity definitions, examples, annotation rules) performs even better and without any fine-tuning. However, the models used are more recent than the backbone model of GoLLIE (CodeLLaMa), so their capabilities have greatly improved since then. Surprisingly, we observe that Gemma 4 obtains the best F1 score overall, even beating GPT-5.6 Sol. The task remains difficult overall (partly due to multiple challenges such as discontinuous entities, which are not handled by the methods we used).

5.3 Entity-type-level results

To better understand where the overall performance gains and losses originate, we report entity-type-level results for the best-performing model on FoodClimNER: `Gemma-4-31B-it` (Table 4).

For the different food practice entity types, we globally observe reasonable performances given the zero-shot setting and corpus difficulty (between 31 and 51 F1 points), meaning our expert guidelines given to the model are useful, as verified in our ablation experiment (Section 6). For the climate change entity types, we notice greenhouse gases are easy to detect (lexically simple as there are not a lot of different possibilities), while impacts and problem origins obtain low performance (around 6 F1 points) showing that these types are harder to detect (very low recall for both) and might benefit from additional examples and refined definitions.

6 Ablation Study

We perform a basic ablation study on our custom prompt method, showing how each part contributes to the final performance by doing 4 runs, removing each time a different part of the prompt. We perform this ablation only on our best performing model, Gemma 4. The results are presented in Table 5. For the "without persona" experiment, instead of stating a NER-specific persona, we just give a neutral persona.

The ablation results show that every component of the prompt contributes positively to the performance, since removing any single element degrades the F1 relative to the full configuration.

Model	Partial			Strict		
	P	R	F1	P	R	F1
Baselines						
GLiNER	17.83	34.36	23.48 [21.21, 25.89]	6.88	13.26	9.06 [7.30, 10.90]
GoLLIE-7B	50.64	25.07	33.53 [30.65, 36.53]	25.32	12.53	16.77 [14.25, 19.39]
GoLLIE-34B	47.91	26.65	34.25 [30.57, 37.98]	16.98	9.44	12.14 [9.61, 14.79]
Prompting						
Gemma4	**62.25**	**49.70**	**55.27 [51.97, 58.44]***	**42.33**	**33.79**	**37.58 [33.89, 41.21]***
Qwen3.5-9B	48.10	32.01	38.44 [35.13, 41.71]	29.54	19.66	23.61 [20.31, 26.85]
Qwen3.5-27B	56.23	46.80	51.08 [47.92, 54.05]	36.51	30.39	33.17 [30.00, 36.34]
GPT5.6-Sol	56.00	40.40	46.94 [43.19, 50.63]	37.62	27.14	31.53 [28.01, 35.04]

Table 3: Models performance comparison on FoodClimNER (Partial, Strict). Best per metric in bold, second-best underlined (descriptive). F1 shown as point estimate [95% document-level bootstrap confidence interval (CI)]. * = significantly better than all others (paired bootstrap, B=10000)

Entity Type	Partial			Strict		
	P	R	F1	P	R	F1
Food Consumption Uses	23.81	45.45	31.25	23.81	45.45	31.25
Food Preparation	86.00	65.15	74.14	55.00	41.67	47.41
Food Repertoire	75.00	57.58	65.14	58.77	45.12	51.05
Food Logistics Travel	87.50	58.33	70.00	50.00	33.33	40.00
Food Storage	41.38	60.00	48.98	29.31	42.50	34.69
Food Supply Purchases	52.94	44.55	48.39	36.47	30.69	33.33
Food Waste Management	59.26	66.67	62.75	41.98	47.22	44.44
Climate Greenhouse Gases	71.43	83.33	76.92	57.14	66.67	61.54
Climate Impacts	33.33	21.05	25.81	8.33	5.26	6.45
Climate Mitigations	64.10	45.05	52.91	43.59	30.63	35.98
Climate Problem Origins	63.64	5.15	9.52	45.45	3.68	6.80
Environmental Concern	25.51	67.57	37.04	17.35	45.95	25.19

Table 4: Entity-type performance on FoodClimNER (Partial, Strict) for gemma-4-it-31-B.

Model	P	R	F1
Full prompt	43.47	34.54	38.49
w/o persona	42.26	33.99	37.68
w/o definitions	43.68	30.21	35.71
w/o rules	39.49	32.37	35.58
w/o examples	34.70	28.22	31.13
w/o all	15.71	16.86	16.27

Table 5: Ablation performances of our custom prompt for gemma-4-31B-it on FoodClimNER (strict evaluation only).

The most critical component is the set of examples: removing them causes the largest drop, suggesting that few-shot demonstrations are essential for the model to correctly delimit and classify entities. Removing the rules and definitions each cost around 2.8-2.9 points, indicating that explicit entity definitions and annotation guidelines meaningfully help disambiguate borderline cases but are less decisive than concrete examples. The persona component has the smallest impact, implying that framing the model's role contributes only marginally to the extraction quality compared to the more content-bearing parts of the prompt (examples, definitions, rules). Removing all parts goes back to giving only the entity type labels to the model, which drastically reduces its capacity (losing half of its maximum points for F1-score).

7 Conclusion

We introduced FoodClimNER, a manually annotated NER corpus extracted from 10 open-access food science and climate change articles, together with expert-crafted annotation guidelines covering seven food practice subtypes and six climate change subtypes, including nested and discontinuous mentions. We evaluated two existing base-

lines (GLiNER, GoLLIE) and a custom zero-shot prompting method relying on explicit entity type definitions and examples, using several recent open-weight and proprietary LLMs.

Our results show that providing full-text definitions, rather than only entity type labels, substantially helps zero-shot LLMs on this task: prompting-based approaches outperform both baselines despite using no task-specific fine-tuning. However, the task remains far from solved. Even the best model, Gemma 4 31B-it, only reaches a strict F1 of 38.49 on the full schema. Climate problem origins in particular prove considerably harder to recognize than food practice subtypes, independently of their frequency in the corpus. This confirms that our expert-defined entity types constitute a genuinely challenging testbed for current zero-shot NER methods and suggests that gains on more generic NER benchmarks do not necessarily transfer to expert-defined, domain-specific schemas such as ours. Although strict span-level evaluation is challenging, the predictions remain useful for assisting experts in identifying relevant text segments for manual review.

We hope FoodClimNER can serve both as a resource for evaluating future NER methods on expert-defined, fine-grained schemas with nested and discontinuous mentions and as a stepping stone toward downstream applications aiming to detect emerging, climate change-driven food practices that could represent future microbiological risks.

8 Limitations

Annotator agreement As we have only one expert annotating the corpus, we cannot compute an agreement for the different entities, and the annotation is subject to bias. However, our annotator has multiple expertise so having another annotator with a similar expertise was difficult to find in the limited time we had to develop the corpus.

Discontinuous entities The methods we use do not handle the discontinuous entities, but they constitute only a small part of the evaluation corpus (8 %).

Custom Prompt The prompt we developed has all the expert-defined annotation rules, but we did not try to optimize it more than this; it could probably be more detailed. Moreover, we did not try to change the output prompt format, which can also influence the performance of the model.

Size of our corpus The size of the corpus is limited because expert annotation is expensive. So we use it only to evaluate zero-shot methods, as there is not enough data to train a model. However, this is only a first version of the corpus, and we plan to extend it with additional documents, more annotations, and better representation of currently underrepresented or absent entity types, potentially including other text sources such as press articles and social media.

9 Acknowledgments

This work was funded by the French national research agency ANR under the reference ANR-24-CE21-0244. It was also provided with computing and storage resources by GENCI at IDRIS thanks to the grant 2026-AD010317377 on the supercomputer Jean-Zay's A100 partition .

References

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528, Online. Association for Computational Linguistics.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Mike Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain Costes, Sergi Blanco-Cuaresma, and 17 others. 2024. INDUS: Effective and Efficient Language Models for Scientific Applications. *Preprint*, arXiv:2405.10725.

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, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ruichu Cai, Junhao Lu, Zhongjie Chen, Boyan Xu, and Zhifeng Hao. 2025. Handling Missing Entities in Zero-Shot Named Entity Recognition: Integrated Recall and Retrieval Augmentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume*

1: Long Papers), pages 10790–10802, Albuquerque, New Mexico. Association for Computational Linguistics.

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. 2025. ZeroNER: Fueling Zero-Shot Named Entity Recognition via Entity Type Descriptions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15594–15616, Vienna, Austria. Association for Computational Linguistics.

Dominique Desjeux. 1998. Les échelles d'observation de la consommation. *Comprendre le consommateur. Paris: Éditions Sciences Humaines*.

Jessica Dietrich, Jens-Andre Hammerl, Annette Johne, Oliver Kappenstein, Christopher Loeffler, Karsten Nöckler, Bettina Rosner, Astrid Spielmeyer, Istvan Szabo, and Martin H Richter. 2023. Impact of climate change on foodborne infections and intoxications. *Journal of health monitoring*, 8(Suppl 3):78.

Ramona A. Duchenne-Moutien and Hudaa Neetoo. 2021. Climate change and emerging food safety issues: A review. *Journal of Food Protection*, 84(11):1884–1897.

EFSA. 2010. Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA journal*, 8(6):1637.

Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Riste Stojanov, Jan Drole, Gordana Ispirova, Giulia Menichetti, Nives Ogrinc, Dimitar Trajanov, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2026. Large language models in food and nutrition science: Opportunities, challenges, and the case of FoodyLLM. *Current Research in Food Science*, 12:101351.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Kristie Hadley, Stefan Wheat, Heidi Honegger Rogers, Arjun Balakumar, Diana Gonzales-Pacheco, Sarah Shrum Davis, Hanna Linstadt, Tracy Cushing, Lewis H Ziska, Christi Piper, and 1 others. 2023. Mechanisms underlying food insecurity in the aftermath of climate-related shocks: a systematic review. *The Lancet Planetary Health*, 7(3):e242–e250.

Morgane Innocent, Agnes Francois Lecompte, Samuel Guillemot, and Ronan Divard. 2024. Environmentally sustainable food: an analysis of consumer practices and their diffusion in france. *British Food Journal*, 126(2):548–577.

Intergovernmental Panel on Climate Change (IPCC). 2023. *Food, Fibre and Other Ecosystem Products*, page 713–906. Cambridge University Press.

Gyula Kasza, Nina Veflen, Joachim Scholderer, Lars Münter, László Fekete, Eszter Zita Csenki, Annamária Dorkó, Dávid Szakos, and Tekla Izsó. 2022. Conflicting issues of sustainable consumption and food safety: Risky consumer behaviors in reducing food waste and plastic packaging. *Foods*, 11(21).

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Yann Le Bodo, Romaric Caillet, and Estelle Chaix. 2026. Mapping Climate-Related Food Practices: A Data-Driven Framework to Anticipate Microbiological Risks. FoodMicro2026. Poster.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. *Preprint*, arXiv:2307.03172.

Marina Arribas Lopez, Yamine Bouzembrak, and Bedir Tekinerdogan. 2026. Artificial Intelligence in Food Safety: A Tertiary Study. *Comprehensive Reviews in Food Science and Food Safety*, 25(2):e70443.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37 of *AAAI'23/IAAI'23/EAAI'23*, pages 13318–13326. AAAI Press.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Huitong Pan, Mustapha Adamu, Qi Zhang, Eduard Dragut, and Longin Jan Latecki. 2025. ClimateIE: A Dataset for Climate Science Information Extraction. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change*

(ClimateNLP 2025), pages 76–98, Vienna, Austria. Association for Computational Linguistics.

Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2019. FoodBase corpus: A new resource of annotated food entities. *Database: The Journal of Biological Databases and Curation*, 2019:baz121.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. *Preprint*, arXiv:2310.03668.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, A. J. Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. OpenAI GPT-5 System Card. *Preprint*, arXiv:2601.03267.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, and 282 others. 2026. Gemma 4 Technical Report. *Preprint*, arXiv:2607.02770.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Sonja J. Vermeulen, Bruce M. Campbell, and John S.I. Ingram. 2012. Climate change and food systems. *Annual Review of Environment and Resources*, 37(Volume 37, 2012):195–222.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. *Preprint*, arXiv:2304.08085.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Preprint*, arXiv:2201.11903.

Ania Wróblewska, Agnieszka Kaliska, Maciej Pawłowski, Dawid Wiśniewski, Witold Sosnowski, and Agnieszka Ławrynowicz. 2022. TASTEset – Recipe Dataset and Food Entities Recognition Benchmark. *Preprint*, arXiv:2204.07775.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical Study of Zero-Shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: Schema-Driven Multi-Task Learning for Structured Information Extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 130–140, Suzhou, China. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. *Preprint*, arXiv:2311.08526.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *Preprint*, arXiv:2308.03279.

A Annotation guidelines

In this section, we specify, in addition to section 3.2, the rules created by our annotator to help annotate the defined entity types in some specific and ambiguous cases. First, we define rules for specific types of entities, starting with food practices entities:

- For mentions of "behavior" related to food or diet, it should be annotated as "Food Repertoire" only if the text links these behaviors explicitly to climate change.
- Food practices must explicitly or implicitly involve a consumer. Macro-level phenomena, or those occurring downstream in the food chain (e.g., primary production, industrial processes), are not considered practices. The consumer must be involved directly when a food practice is annotated.
- Do not annotate mentions where we cannot identify a clear practice (e.g., consumer habits); too generic a term is not informative.

- Do not annotate food items (e.g., meat dishes, sustainable food) but take care of similar forms that refer to repertoire and that must be annotated (e.g., meat-free dishes).
- Food categories (e.g., wild foods, edible insects, organic foods) are not annotated as food practices unless they are embedded in a consumption or dietary context (e.g., consumption, diet, intake, eating patterns).
- Cognitive constructs are excluded unless they directly refer to operationalized food preparation activities or competencies immediately applied in food acquisition, preparation, or consumption.
- Include modifiers expressing intensity or direction when a practice is part of the observed behavioral change (e.g., increase/decrease).
- A practice that includes negation is still a practice and should be annotated (e.g., "Do not rinse meat", "Do not pay attention to expiration dates").

Then we also have rules for climate change entities :

- If a climate hazard is named in more detail in a text, the entire term is annotated (e.g., surface water flood or soil liquefaction).
- For the climate impact entity, vector-borne diseases and other diseases explicitly described as climate-sensitive are annotated as impacts. COVID-19 is annotated as an impact only when the text explicitly establishes a link between climate change and the pandemic.
- If an increase or decrease in activity reduces climate change, it should be annotated as a Climate Mitigation (e.g., "food waste has seen a notable decrease in volume").
- Transitions, strategies, policies, or transformations explicitly aimed at the sustainability of a system should be annotated as climate mitigation even when climate change is not explicitly mentioned.
- Environmental concern includes biodiversity, the ecological footprint, etc. Environmental concern includes generic or non-actionable references to sustainability, environmental awareness, or trends in consumption that do not describe explicit consumer actions.

A.1 Global rules

- Discontinuous annotation: If a practice is expressed as a predicate or attribute of diets (e.g., be varied, be seasonal), annotate the full conceptual unit of dietary recommendation using discontinuous spans including the implicit subject (e.g., diets should), and include all coordinated attributes as separate linked spans.
- Nested annotation: a food practice can have the same exact boundary as another entity type. For example, as some practices may have a mitigating effect on climate change, they can also be a climate mitigation entity. In the case of discontinuous mentions, we might even have some spans that have the same types for two different discontinuous mentions (but the other spans composing these mentions will be different).

B Simplified corpus results

As mentioned in section 3.2, we evaluate on two versions of our corpus, FoodClimNER (containing all entity types) and FoodClimNER-simple (aggregating all food practice types into a single type and all climate change types in a single type as well). In this section we provide the results of FoodClimNER-simple in Tables 6 and 7.

Model	P	R	F1
Baselines			
GLiNER	33.69	26.06	29.38
GoLLIE-7B	35.39	15.23	21.30
GoLLIE-34B	38.82	17.10	23.74
Custom Prompt			
Gemma 4 31B-it	**57.91**	<u>32.29</u>	<u>41.47</u>
Qwen 3.5-9B	54.91	18.99	28.22
Qwen 3.5-27B	<u>55.67</u>	25.69	35.15
GPT-5.6 Sol	51.35	**38.44**	**43.97**

Table 6: Strict micro-averaged metrics on the FoodClimNER-simple corpus. Best per metric in bold, second-best underlined.

This time, the results are different from those on the harder version of the corpus. First, the GLiNER model performs better than the GoLLIE models. entity type labels are already sufficient information because they are easier to apprehend than the more fine-grained types.

Globally, the models achieve better performance with these aggregated entity types. We also observe that GPT-5.6 obtains the best results this time.

Our custom prompt with more recent models still beats all the baselines (except for the Qwen 3.5 9B parameter model, which performs slightly below GLiNER).

Similarly to the fine-grained type evaluation, performance is higher for food practice entities than for climate change entities (Table 7).

Entity Type	P	R	F1
Food Practice	56.22	43.98	49.35
Climate Change	41.88	28.93	34.22

Table 7: Strict entity-type performances on FoodClimNER-simple (Partial, Strict) for gpt-5.6-sol.

C Compute cost and environmental cost

We measure the number of GPU hours spent on our experiments and the estimated environmental cost on our GPU cluster. We spent 135[6] GPU hours (on NVIDIA A100 80 GB GPUs) running all the open-weight models on the two versions of our corpus (FoodClimNER and FoodClimNER-simple) which translates to 8.5 kg CO2e according to the estimation given by our GPU cluster. We spent $12 for running the experiment with the OpenRouter API (for the GPT-5.6 Sol model). However, we do not have any way to compute an environmental cost for this model.

D Complete Custom Prompt

In Figure 3, we provide the full template filled with our corpus entity type labels, definitions, examples, and annotation rules.

[6]This is the value given by our cluster in normalized V100 equivalent hours, because they attribute GPU hours budgets depending on the hardware we ask (higher rate for better GPUs). So for A100 GPUs, hours are counted double so it is effectively 67.5 A100 hours.

System prompt

```
You are a helpful assistant specialized in named entity recognition.
```

User prompt

```
Extract all mentions of the declared entities types from the input text in the specified format.

### Entity Types

* climate_greenhouse_gases
  - Definition: Gases that cause heating of the atmosphere.
  - Examples: carbon dioxide, CH4, GHG emission
* climate_hazards
  - Definition: Hazards with potential negative impact on climate. Where a hazard is named in more detail in a text, the
      entire term is annotated.
  - Examples: floods, wildfires, droughts, heatwaves, surface water flood, soil liquefaction
* climate_impacts
  - Definition: Effects of hazards, primarily negative effects on humans. Climate impacts on livestock are also
      considered impacts, as they indirectly affect humans. COVID-19 is annotated as impact only when the text explicitly
       establishes a link between climate change and the pandemic.
  - Examples: Vector-borne diseases, diseases explicitly described as climate-sensitive
* climate_mitigations
  - Definition: Activities to reduce climate change or to better deal with the consequences. If an increase or decrease
      in activity reduces climate change, it should be annotated. Transitions, strategies, policies, or transformations
      explicitly aimed at the sustainability of a system should be annotated even when climate change is not explicitly
      mentioned.
  - Examples: food waste reduction, transitions aimed at sustainability, strategies aimed at sustainability, policies
      aimed at sustainability
* climate_problem_origins
  - Definition: Problems that describe why the climate is changing. Sectors cited as causes of energy use are also
      annotated as problem origins.
  - Examples: fossil fuel, deforestation, transport sector
* environmental_concern
  - Definition: An intermediate/residual category within the environmental sphere for which the link to climate is not
      established or direct. It does not specifically refer to climate change. Includes generic or non-actionable
      references to sustainability, environmental awareness, or trends in consumption that do not describe explicit
      consumer actions.
  - Examples: biodiversity, ecological footprint, pollution, water
* food_consumption_uses
  - Definition: Covers consumption practices related to eating and meals.
  - Examples: portion sizes, choice of utensils
* food_culinary_preparation
  - Definition: All attitudes and practices related to food preparation and cooking. It also includes skills and actions
      that we perform when cooking.
  - Examples: raw food diets, water conservation during cooking, reducing energy consumption in cooking, minimizing waste
       during preparation
* food_repertoire
  - Definition: Covers practices and factors related to food choices (in terms of quality or quantity) that may or may
      not be part of an individual's diet. Mentions of behaviour related to food or diet should be annotated only if the
      text links these behaviours explicitly to climate change.
  - Examples: meat consumption, food choices
* food_logistics_travel
  - Definition: Includes pre-purchase practices and travel to the place of purchase.
  - Examples: checking existing resources before shopping, meal planning, using low-carbon transportation to go grocery
      shopping
* food_storage
  - Definition: Covers food storage practices after purchase and before preparation.
  - Examples: choice of packaging, adherence to expiration dates, disregard of expiration dates, preservation techniques
* food_supply_purchases
  - Definition: Covers purchasing practices (product selection, choice of retail locations, interactions with marketing)
      as well as non-market procurement strategies.
  - Examples: product selection, choice of retail locations, interactions with marketing, self-production
* food_waste_management
  - Definition: Covers practices that take place after a meal, with consideration for the end of a food item's life cycle.

  - Examples: storing leftovers, composting, sorting food waste, expiration date reasoning

### Global Annotation Conventions

- Discontinuous annotation: if a practice is expressed as a predicate or attribute of diets (e.g. be varied, be seasonal),
       annotate the full conceptual unit of dietary recommendation using discontinuous spans including the implicit
      subject (e.g. diets should), and include all coordinated attributes as separate linked spans.
- Nested annotation: a food practice can have an exact boundary as another entity type. For example, as some practices
      may have a mitigating effect on climate change they can also be a climate mitigation entity. In the case of
      discontinuous mentions, we might have some spans that have the same types for two different discontinuous mentions
      (but the other spans composing these mentions will be different).

### Food Practices General Guidelines
```

```
- Food practices must explicitly or implicitly involve a consumer. Macro-level phenomena, or those occurring downstream
      in the food chain (e.g., primary production, industrial processes), are not considered practices. The consumer must
       be involved directly when a food practice is annotated.
- Do not annotate mentions where we cannot identify a clear practice (e.g. consumer habits) -- a too generic term is not
      informative.
- Do not annotate food items (e.g. meat dish, sustainable food) but take care of similar forms that refer to repertoire
      and that must be annotated (e.g. meat-free dishes).
- Food categories (e.g., wild foods, edible insects, organic foods) are not annotated as food practices unless they are
      embedded in a consumption or dietary context (e.g. consumption, diet, intake, eating patterns).
- Cognitive constructs are excluded unless they directly refer to operationalized food preparation activities or
      competencies immediately applied in food acquisition, preparation, or consumption.
- Include modifiers expressing intensity or direction when a practice is part of the observed behavioral change (e.g.,
      increase/decrease).
- A practice that includes negation is still a practice and should be annotated (e.g. "Do not rinse meat", "not pay
      attention to expiration dates").

### Output Format

Return a JSON object where each key is an entity type and the value is a list of all mentions of that entity type found
      in the text.

Example:
“‘json


{
  "entity_type1" : [
    "mention_type1_1",
    "mention_type1_2",
    "..."
  ],
  "entity_type2" : [
    "mention_type2_1",
    "..."
  ],
  "..."
}

“‘

Return only valid JSON, no additional text.

### Input Text

“‘txt
{input_text}
“‘
```

Figure 3: Prompt used for LLM-based named entity recognition on the FoodClimNER corpus.

Fine-Tuned Encoders versus Zero-Shot Prompting for Detecting Threat and Solution Framing in German Climate News

Raven Adam[1] **David Maier**[1] **Marie Kogler**[1,2]
[1]University of Graz, Austria [2]Graz University of Technology, Austria

Abstract

News media play a central role in shaping public perceptions of climate change, and whether coverage emphasizes threats or solutions affects audience engagement and policy support. We compare fine-tuned encoders and zero-shot prompting for classifying sentences from German-language climate news as threat-oriented, solution-oriented, both, or neither, using two independent binary classifiers. We fine-tune three German encoders with context windows ranging from the target sentence alone to four preceding and four following sentences, and compare them with three open-weights generative models and the non-generative decision model Jev. All models are evaluated with article-grouped five-fold cross-validation on 440 manually coded Austrian newspaper articles. The best encoder, GBERT-large with the widest context, achieves weighted F1 scores of 0.805 (threat) and 0.802 (solution). The recently released MiMo-V2.6-Flash reaches 0.807 and 0.798 without task-specific training, placing both approaches on par, whereas earlier prompting models remain below the best encoder. Additional context improves all encoders consistently, by up to 4.8 F1 points. On the natural class distribution, MiMo-V2.6-Flash outperforms GBERT-large on threat classification, while GBERT-large retains a small advantage for solution classification. These results indicate that the gap between fine-tuned encoders and prompted language models has only recently been closed.

Keywords: climate change framing, text classification, BERT, large language models, zero-shot prompting, German NLP, computational content analysis

1 Introduction

How the media frames climate change affects what audiences understand, feel, and are willing to do about it. Research in climate communication has repeatedly shown that threat-oriented framing, which highlights risks and dangers, and solution-oriented framing, which presents actions and remedies, produce different audience responses (Hart and Feldman, 2014; Feldman and Hart, 2018; Ringhofer et al., 2026). Threat framing without accompanying efficacy information can induce fatalism, while solution-oriented coverage promotes constructive engagement and policy support (Feldman et al., 2015; Thier and Lin, 2022). Tracking the balance between these two framing types across large media corpora is therefore of direct interest to communication researchers studying climate discourse.

Content analysis remains the primary method for studying media framing, but manual coding is labor-intensive and difficult to scale. A single study may require weeks of trained human effort to code a few hundred articles (Grimmer and Stewart, 2013; Boumans and Trilling, 2016). Automated approaches based on natural language processing (NLP) can process orders of magnitude more text, though their reliability must be validated against human judgments (Grimmer and Stewart, 2013). Among the available automated approaches, two have become dominant in recent years. The first involves fine-tuning pre-trained transformer-based encoder models such as BERT (Devlin et al., 2019) on labeled training data. The second relies on prompting large language models (LLMs) to perform classification through in-context learning, without parameter updates (Brown et al., 2020; Gilardi et al., 2023).

Both approaches have demonstrated strong performance on various text classification tasks, but they involve fundamentally different trade-offs. Fine-tuned models require labeled training data and computational resources for training, but produce fast and deterministic inference. LLM-based approaches require no training data beyond a few optional examples in the prompt, but depend on careful prompt design and incur higher per-sample inference costs. Recent comparative studies find

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 158–169
October 28, 2026 ©2026 Association for Computational Linguistics

that fine-tuned encoder models often outperform zero-shot and few-shot LLMs on classification tasks, while prompting-based approaches such as chain-of-thought and structured outputs can improve performance in some settings (Bucher and Martini, 2024; Galke and Scherp, 2024; Edwards and Camacho-Collados, 2024; Wei et al., 2022; Dunivin, 2025).

Despite growing interest in automated framing detection (Ali and Hassan, 2022; Piskorski et al., 2023) and LLM-based content analysis (Gilardi et al., 2023; Ziems et al., 2024), no study has directly compared these two approaches for detecting climate-specific framing at the sentence level. Most computational framing work uses generic frame taxonomies (Card et al., 2015) rather than domain-specific categories. Generic frames, however, do not capture the threat-versus-solution distinction that is central to climate communication research, and classifiers trained on one frame taxonomy do not transfer reliably to another. Similarly, work on German-language climate media remains limited (Hase et al., 2021; Adam et al., 2023; Adam and Kogler, 2025), even though framing patterns, editorial conventions, and linguistic structures differ across languages in ways that affect classifier performance (Piskorski et al., 2023; Mondshine et al., 2025).

This paper makes four contributions. First, we present a systematic comparison of three fine-tuned German encoder models and four zero-shot prompting models, including a non-generative decision model, for sentence-level climate framing classification, evaluated with article-grouped cross-validation on manually coded German newspaper articles. Second, we show that the most recent prompting model matches the best fine-tuned encoder on balanced evaluation data and partly outperforms it under the natural class distribution, whereas earlier prompting models remain below the encoder, indicating that the performance gap between the two paradigms has only recently been closed. Third, we quantify the effect of context on encoder-based classification by comparing four context windows, demonstrating consistent improvements of up to 4.8 F1 points over single-sentence classification. Fourth, we describe a prompt architecture for zero-shot framing classification that incorporates chain-of-thought reasoning, a speech-act taxonomy, and domain-specific coding rules in the source language.

2 Related Work

2.1 Threat and Solution Framing in Climate Communication

Framing, as defined by Entman (1993), involves selecting and making salient certain aspects of a perceived reality to promote particular problem definitions, causal interpretations, and treatment recommendations. In climate communication, framing research has identified a wide range of recurring patterns. Nisbet (2009) proposed a typology of climate-specific frames including scientific uncertainty, economic development, and morality. Badullovich et al. (2020) mapped 274 studies and found scientific, economic, and environmental frames to be dominant, while Guenther et al. (2024) identified 18 pre-defined frames across 25 years of literature, with "climate action" and "harmful impacts" as the two most commonly studied.

The distinction between threat and solution framing is particularly relevant for understanding audience responses. Hart and Feldman (2014) found that U.S. television news frequently conveys climate threats but provides inconsistent efficacy information, meaning that impacts and potential actions rarely appear together. Feldman et al. (2015) extended this finding to print media and documented partisan differences in how newspapers balance threat and efficacy. More recently, Guenther et al. (2022) analyzed the shift from doom framing toward sustainable solutions in international news magazines, and Thier and Wu (2024) conducted the most detailed empirical study of solution-oriented climate framing to date, analyzing 244 articles and identifying three previously undescribed solution frames.

Research on German-language and Austrian climate media remains limited but growing. Hase et al. (2021) found that climate coverage in Germany accounted for only 0.24% of all articles in their ten-country sample. Adam et al. (2023) analyzed long-term trends in Austrian newspaper coverage using NLP pipelines. More recently, Dablander et al. (2025) used GPT-4-assisted analysis to study German media coverage of climate activist groups, and Meyer et al. (2025) examined framing differences across 21 German outlets with varying political orientations.

2.2 Transformer-Based Text Classification

The Transformer architecture (Vaswani et al., 2017) and its bidirectional variant BERT (Devlin et al.,

2019) established the current paradigm for text classification through pre-training and fine-tuning. BERT's sentence-pair input format, which concatenates two text segments with special separator tokens, was originally designed for tasks such as natural language inference and paraphrase detection. This format also enables classification of a target sentence with additional contextual information from a second segment.

For German-language tasks, Chan et al. (2020) introduced GBERT and GELECTRA, monolingual German models trained with whole-word masking. These models outperform multilingual alternatives such as XLM-RoBERTa (Conneau et al., 2020) on German NLP benchmarks (Chan et al., 2020; Wunderle et al., 2025), making monolingual models the preferred choice for single-language classification tasks.

The use of surrounding context for sentence-level classification has been explored in several studies. Cohan et al. (2019) showed that incorporating document context through joint sentence representations improves sequential sentence classification. Luoma and Pyysalo (2020) studied the effect of cross-sentence context for named entity recognition with BERT, finding consistent improvements when target sentences are placed within their document context. These findings motivate the context configurations used for our fine-tuned encoders.

In the climate domain specifically, Webersinke et al. (2022) demonstrated that domain-adaptive pre-training on climate text (ClimateBERT) improves downstream task performance, and Stammbach et al. (2023) applied fine-tuned transformers to sentence-level classification of environmental claims. For frame detection, Liu et al. (2019a) achieved 84% accuracy by fine-tuning BERT on news headlines, and the SemEval-2023 shared task on multilingual framing detection (Piskorski et al., 2023) saw top systems based on fine-tuned XLM-RoBERTa.

2.3 LLM-Based Text Classification

Brown et al. (2020) demonstrated that large language models can perform NLP tasks through in-context learning with a small number of examples in the prompt. Wei et al. (2022) showed that chain-of-thought (CoT) prompting, where the model is prompted to produce intermediate reasoning steps before a final answer, substantially improves performance on complex tasks. These two techniques form the basis of most current approaches to LLM-based classification.

Several studies have evaluated LLMs specifically as text annotation tools for social science research. Gilardi et al. (2023) found that zero-shot ChatGPT outperforms crowd workers on frame detection and other annotation tasks. Alizadeh et al. (2025) extended this work to open-source LLaMA- and Mistral-based models, showing that fine-tuning can narrow the gap to proprietary systems. Dunivin (2025) demonstrated that CoT prompting substantially improves intercoder reliability for qualitative coding tasks, and Farjam et al. (2025) presented a practical workflow for LLM-assisted content analysis in communication research.

Prompt design has become a research area in its own right. Schulhoff et al. (2024) catalogued 58 distinct prompting techniques, while Mei et al. (2025) framed context engineering as a broader discipline encompassing retrieval, processing, and management of all contextual information provided to LLMs during inference. For multilingual applications, Mondshine et al. (2025) showed that the relative benefit of using source-language prompts versus various translation strategies is highly dependent on both the task and the model, meaning that the optimal prompt language cannot be assumed in advance and must be evaluated empirically.

3 Data

3.1 Corpus

The corpus consists of 440 articles from two Austrian daily newspapers: *Kronen Zeitung* (304 articles) and *Der Standard* (136 articles), published between January and June 2024. These newspapers were selected as the two highest-circulation Austrian dailies excluding regional and free publications, and they represent contrasting editorial approaches. Articles were retrieved from the WISO database (GBI-Genios) using the search string `klima* OR CO2`, which was determined through iterative refinement to capture the same relevant articles as a broader set of climate-related keywords while producing fewer irrelevant results (Adam et al., 2023). Retrieved articles were manually screened for thematic relevance to climate change or climate protection, excluding article types such as letters to the editor and announcements. Articles that did not meet the criteria upon closer inspection during coding were removed, yielding the final corpus of 440 articles. A de-

tailed description of the corpus, the coding scheme, and the annotation process is provided by Riesinger (2024).

3.2 Annotation

The articles were coded manually using MAXQDA software following a detailed coding scheme developed in collaboration with linguists as part of a broader interdisciplinary research project. The coding scheme distinguishes two primary categories: *threat orientation* and *solution orientation*. These are treated as independent dimensions rather than mutually exclusive classes, meaning that a text segment can be coded as threat-oriented, solution-oriented, both, or neither.

Threat orientation is defined as the communication of risks and dangers in the context of the climate crisis, drawing on Entman's (1993) framing concept and Searle's taxonomy of speech acts. Threat-oriented statements typically function as commissives (promising, threatening) or directives (requesting, recommending) that aim to prompt action. Solution orientation is defined following similar linguistic principles and encompasses directives and assertives that describe, recommend, or evaluate actions to address climate change.

To ensure coding reliability, categories were precisely defined and iteratively refined over a five-week period. For approximately 8% of the material, two coders independently assigned the categories used in the quantitative analysis. This process was conducted in multiple cycles with interim discussions and adjustments until a high level of agreement was reached. Ambiguous cases were resolved through discussion with additional annotators, including the linguists involved in the project.

3.3 Sentence Segmentation and Dataset

The manually coded segments were split into individual sentences using the German HDT model from the Stanza NLP toolkit, which provides state-of-the-art tokenization and sentence segmentation for German text. Each sentence inherits the binary labels of its parent segment: a sentence is labeled as threat-oriented if the segment it belongs to was coded as threat-oriented, and analogously for solution orientation.

Sentence segmentation yielded 10,981 sentences from the 440 articles, of which 2,575 were labeled as threat-oriented and 4,108 as solution-oriented. Since a sentence can carry both labels, these counts are not mutually exclusive. The remaining sentences carry neither label.

Because the positive class is substantially smaller than the neutral class for both tasks, we balanced each binary dataset by randomly sampling neutral sentences to match the number of positive sentences, yielding 5,150 sentences for the threat task and 8,216 sentences for the solution task.

Rather than relying on a single train-test split, we evaluate all models using five-fold cross-validation with stratified group partitioning (`StratifiedGroupKFold` from scikit-learn). Folds are grouped by article, so that all sentences from a given article are assigned to the same fold and no article contributes sentences to both the training and the test set within any fold. This prevents topical and stylistic overlap between training and test data that could otherwise inflate performance estimates. Within this constraint, folds are stratified by the binary label to keep the class distribution similar across folds. Because entire articles are assigned to folds, fold sizes vary slightly, and each fold corresponds to an approximately 80/20 split between training and test data. The fold assignments were computed once per task and stored with the dataset, so that every model is trained and evaluated on identical folds.

4 Methods

4.1 Task Formulation

We formulate the framing classification as two independent binary classification tasks: one for threat orientation and one for solution orientation. Each classifier produces a binary decision (positive or negative) for a single target sentence, though both methods additionally receive contextual information as described in Sections 4.2 and 4.3. This decomposition into two binary tasks rather than a single four-class task reflects the structure of the underlying coding scheme, where threat and solution are independent dimensions. A sentence can be classified as both threat- and solution-oriented if both binary classifiers return a positive prediction, or as neither if both return negative predictions.

4.2 Fine-Tuned Encoder Classification

We fine-tune three pre-trained German encoder models of different architectures and sizes. `deepset/gbert-large` (Chan et al., 2020) is a monolingual German BERT model with 335M parameters, pre-trained with whole-

word masking. `deepset/gelectra-large` (Chan et al., 2020) is a German ELECTRA model of the same size, pre-trained with the replaced-token-detection objective (Clark et al., 2020). `GeistBERT/GeistBERT_base` (Scheible-Schmitt and Frei, 2025) is a smaller German model following the RoBERTa base configuration (Liu et al., 2019b), pre-trained with whole-word masking. For all three models, we use the standard HuggingFace sequence classification architecture (Wolf et al., 2020), which places a classification head on top of the final-layer representation of the first input token.

To incorporate contextual information, we use the models' native two-segment input format, with the target sentence in the first segment and its surrounding context in the second:

`[CLS]` *target sentence* `[SEP]` *context* `[SEP]`

For GeistBERT, the model-specific equivalents of these special tokens are used. We compare four input configurations that differ in the amount of context provided: (1) the target sentence alone, without a context segment; (2) the target sentence with the immediately preceding sentence; (3) the target sentence with both the immediately preceding and the immediately following sentence; and (4) the target sentence with the four preceding and the four following sentences. When both preceding and following sentences are included, they are separated within the context segment by an additional separator token (*preceding* `[SEP]` *following*). Context sentences that fall outside the article boundaries, for instance preceding sentences for the first sentence of an article, are replaced by empty strings. This input format allows the transformer's self-attention mechanism to attend to contextual cues from the surrounding text when computing the representation used for classification, following the principle that isolated sentences are often ambiguous without their document context (Cohan et al., 2019; Luoma and Pyysalo, 2020).

For each combination of model and input configuration, two separate classifiers are fine-tuned, one for threat classification and one for solution classification. All models are fine-tuned directly on the task-specific training data without intermediate domain-adaptive pre-training. Training uses the AdamW optimizer (`adamw_torch_fused`) with a learning rate of 2×10^{-5} or 3×10^{-5} and between 10 and 15 epochs, depending on the model and the input configuration.

4.3 Zero-Shot Prompting

The prompting-based approach classifies each sentence in a separate zero-shot inference request, without task-specific training or in-context examples. Each request contains the target sentence together with up to four preceding and four following sentences from the same article, which corresponds to the largest context window used for the fine-tuned encoders (Section 4.2). As with the encoder models, one prompt is designed for threat classification and another for solution classification. The prompts are written entirely in German to match the language of the source texts and the original coding guidelines.

4.3.1 Prompt Architecture

Each prompt is composed of six XML-delimited sections, following established practices in context engineering (Mei et al., 2025; Schulhoff et al., 2024). The prompt opens with a *role definition* establishing the model as a climate journalism analyst, followed by a *speech-act taxonomy* (representative, directive, commissive, expressive, declarative) that provides a linguistic framework for analyzing communicative function. A *framing definition* section then defines the target framing type through guiding questions (who, when, where, why, how) and characteristic linguistic forms, derived from the same coding guidelines used for manual annotation. This is complemented by *classification rules* that operationalize the guidelines into concrete inclusion and exclusion criteria (nine inclusion and three exclusion rules for threat; eight and seven for solution). Finally, the *context sentences* and the *target sentence* to classify are appended. This architecture mirrors the information available to human coders, namely the coding guidelines and the surrounding article text, but deliberately omits labeled examples, so that the models rely solely on the written guidelines.

4.3.2 Structured Output with Chain of Thought

The generative models return a structured JSON object with three fields in a fixed order. The `reason` field appears first, requiring the model to produce a short justification (one to three sentences) before committing to a classification. This ordering follows the chain-of-thought principle (Wei et al., 2022), which has been shown to improve classification accuracy in content analysis tasks (Dunivin, 2025). The `classification` field contains a

Table 1: Evaluated prompting models. Parameter counts for mixture-of-experts (MoE) models are given as total/active.

Model	Params	Arch.	Weights
MiMo-V2.6-Flash	309B/15B	MoE	open
Qwen3.8-27B	27B	dense	open
Llama 4 Maverick	400B/17B	MoE	open
Jev-1.13.0	n/a	decision	closed

boolean value indicating whether the sentence exhibits the framing type. The `confidence` field contains a numeric score reflecting the model's self-assessed certainty. For models that offer a native reasoning mode, this mode was enabled in addition to the `reason` field.

4.3.3 Models

Table 1 summarizes the four evaluated models. `meta-llama/llama-4-maverick` is retained from preliminary work, where it was selected based on cost-performance evaluations by Maier (2025). `xiaomi/mimo-v2.6-flash`[1] is a recently released sparse mixture-of-experts model, and `qwen/qwen3.8-27b` (Qwen Team, 2026) is a dense model of moderate size with a hybrid attention mechanism combining linear and full attention layers. Together, these three open-weights generative models cover both sparse and dense architectures across a range of active parameter counts.

The fourth model, `jev-1.13.0`[2], differs fundamentally from the generative models. Jev is a closed-weights *decision model* that is built on a pretrained language model but trained with reinforcement learning for calibrated decisions (RLCD). Instead of generating free text, it receives structured input state and a question, and returns a typed answer together with a probability. We supplied Jev with the same German-language content as the generative models: the role definition, speech-act taxonomy, framing definition, classification rules, and context sentences were passed as input state, and the classification was posed as a question with a boolean answer type. Because Jev does not generate text, it produces no `reason` field. Jev's primary training language is English, so its application to German input constitutes a cross-lingual setting.

[1] https://mimo.mi.com/models/en-US/mimo-v2.6-flash

[2] https://typesafe.ai/blog/introducing-system-one-models-and-jev

4.3.4 Inference Parameters

The three generative models were accessed through the OpenRouter API, and Jev through its provider's API. Structured output parsing for the generative models was handled by the Vercel AI SDK, which enforces the three-field JSON schema (reason, classification, confidence) at the API level rather than relying on manual extraction from free-form text. All generative inference requests used a temperature of 0 and a fixed random seed (`seed=42`). Each sentence was classified once per model and task in a separate request. Because the prompting approach requires no training, every model classifies each sentence exactly once, and the predictions are grouped by the cross-validation fold of the respective sentence for evaluation.

4.4 Evaluation

All models are evaluated on the same five article-grouped cross-validation folds (Section 3). For the fine-tuned encoders, a separate model is trained on the training split of each fold and evaluated exclusively on the test split of that same fold. For the prompting models, which require no training, the predictions are grouped by the fold of the respective sentence. Our primary metric is the weighted F1 score over both classes on each fold's test split, which accounts for performance on framing and non-framing sentences alike. The weighted F1 score is computed separately for each fold and reported as the mean and standard deviation across the five folds. This comparison covers all encoder models, input configurations, and prompting models.

For the best-performing encoder configuration and the best-performing prompting model, we conduct two additional analyses. First, because the balanced evaluation sets do not reflect the class distribution encountered in practice, we evaluate both models on all 10,981 sentences of the corpus at their natural base rate (23.4% threat-oriented and 37.4% solution-oriented sentences). Because the folds are defined at the article level, each neutral sentence that was excluded by the balanced sampling can be assigned to the test split of the fold containing its source article, where that article is guaranteed not to occur in the training split. Each fold model is then applied to its extended test split, so that across the five folds every sentence of the corpus is classified exactly once without article overlap between training and test data. The prompt-

ing models are applied to the same extended test splits. For this setting, we report positive-class F1 together with weighted and macro F1, pooled over all folds. Second, we isolate the subset of evaluation sentences on which the two models produce different predictions and report precision, recall, and F1 for the positive class, along with overall accuracy, on this disagreement subset.

5 Results

5.1 Main Results

Table 2 reports the weighted F1 scores of all encoder models and input configurations as well as all prompting models, averaged over the five cross-validation folds.

The best fine-tuned encoder and the best prompting model perform on par. GBERT-large with ± 4 context achieves a weighted F1 of 0.805 for threat and 0.802 for solution classification, while MiMo-V2.6-Flash achieves 0.807 and 0.798, respectively. The differences of at most 0.004 are small relative to the standard deviations across folds (0.013–0.022), so neither approach can be considered superior on the balanced evaluation sets.

Among the prompting models, MiMo-V2.6-Flash performs best on both tasks. Qwen3.8-27B and Llama 4 Maverick reach similar scores for threat classification (0.790 and 0.791) and fall slightly further behind on solution classification (0.777 and 0.770). Jev-1.13.0 shows the most uneven profile: its solution classification (0.772) is comparable to that of Qwen3.8-27B and Llama 4 Maverick and has the lowest variance across folds of all models (± 0.006), whereas its threat classification (0.701) is clearly the weakest result in the comparison.

5.2 Effect of Context and Encoder Choice

Adding context improves all three encoders monotonically on both tasks (Table 2). For GBERT-large, weighted F1 increases from 0.760 with the target sentence alone to 0.805 with ± 4 context for threat classification, and from 0.754 to 0.802 for solution classification, gains of 4.5 and 4.8 points. GELECTRA-large shows similar gains (4.3 and 4.4 points), while GeistBERT-base benefits somewhat less (3.0 and 3.8 points). The largest single improvement comes from adding the preceding sentence (2.1 and 2.3 points for GBERT-large), with smaller additional gains from the following sentence (1.3 and 1.4 points) and from extending the window to four sentences on either side (1.2 and 1.1 points). Although the gains between adjacent configurations are of a similar magnitude to the standard deviations across folds, the consistent direction across all models and both tasks indicates a robust effect of context, in line with findings from sequential sentence classification research (Cohan et al., 2019; Luoma and Pyysalo, 2020).

The ranking of the encoders is stable across all configurations and both tasks: GBERT-large performs best, followed closely by GELECTRA-large, while the smaller GeistBERT-base consistently scores lowest. Notably, GBERT-large with only the preceding and following sentence (± 1) already matches Qwen3.8-27B and Llama 4 Maverick, which receive ± 4 context.

5.3 Natural Class Distribution

Table 3 reports the performance of the best encoder (GBERT-large, ± 4) and the best prompting model (MiMo-V2.6-Flash) on all 10,981 sentences at their natural class distribution, pooled over the five folds.

Under the natural distribution, positive-class F1 decreases considerably for both models compared to the balanced setting, most strongly for threat classification, where positive sentences make up less than a quarter of the corpus. For threat classification, MiMo-V2.6-Flash outperforms GBERT-large on all three metrics, by 2.5 points in positive-class F1 and 2.7 points in weighted F1. For solution classification, GBERT-large retains a small advantage of 1.0 point in positive-class F1. For GBERT-large, the decrease is driven by precision: its recall for the positive class remains high (0.819 for threat, 0.833 for solution), but its precision falls to 0.536 and 0.694, respectively, as the model assigns the positive label to a substantial share of the additional neutral sentences.

5.4 Disagreement Analysis

To understand where the two approaches diverge, we isolated the evaluation sentences on which GBERT-large (± 4) and MiMo-V2.6-Flash produce different predictions. The two models disagree on 989 threat sentences (19% of the evaluation set) and 1,519 solution sentences (18%). Table 4 reports the performance of both models on these subsets.

Since the two binary classifiers disagree on every sentence in these subsets, exactly one of them is correct for each sentence, and their accuracies sum to one. Neither model dominates: MiMo-V2.6-Flash is correct on 51.6% of the contested threat

Table 2: Weighted F1 on the balanced test splits (mean ± standard deviation across five article-grouped folds). Context columns: target sentence only (T), target with the preceding sentence (−1), target with the preceding and following sentence (±1), and target with the four preceding and four following sentences (±4). Prompting models receive ±4 context. The best score per task is shown in bold.

	Threat				Solution			
Model	T	−1	±1	±4	T	−1	±1	±4
Fine-tuned encoders								
GBERT-large	.760±.011	.781±.017	.793±.013	.805±.017	.754±.014	.777±.013	.792±.013	**.802**±.013
GELECTRA-large	.759±.013	.775±.016	.793±.018	.802±.019	.750±.022	.773±.013	.784±.014	.794±.017
GeistBERT-base	.752±.026	.763±.016	.775±.015	.782±.019	.743±.013	.765±.014	.776±.010	.781±.012
Zero-shot prompting								
MiMo-V2.6-Flash	–	–	–	**.807**±.014	–	–	–	.798±.022
Qwen3.8-27B	–	–	–	.790±.010	–	–	–	.777±.014
Llama 4 Maverick	–	–	–	.791±.016	–	–	–	.770±.015
Jev-1.13.0	–	–	–	.701±.022	–	–	–	.772±.006

Table 3: Pooled F1 on the full corpus at the natural class distribution (n = 10,981; 23.4% threat, 37.4% solution). Pos. denotes F1 for the positive class.

Task	**Model**	**Pos.**	**Weighted**	**Macro**
Threat	GBERT-large	.648	.804	.750
Threat	MiMo-V2.6-Flash	**.673**	**.831**	**.776**
Solution	GBERT-large	**.757**	**.803**	**.794**
Solution	MiMo-V2.6-Flash	.747	.799	.788

Table 4: Performance on the disagreement subsets, i.e., sentences where GBERT-large (±4) and MiMo-V2.6-Flash produce different predictions. Precision (P), recall (R), and F1 are reported for the positive class; accuracy (Acc) is computed over both classes.

Task	**Model**	**P**	**R**	**F1**	**Acc**	n
Threat	GBERT-large	.510	.612	.557	.484	989
Threat	MiMo-V2.6-Flash	.561	.388	.459	.516	989
Solution	GBERT-large	.467	.583	.518	.513	1,519
Solution	MiMo-V2.6-Flash	.428	.417	.423	.487	1,519

sentences, and GBERT-large on 51.3% of the contested solution sentences. The models differ mainly in their decision tendencies. GBERT-large assigns the positive label more readily and achieves higher recall on both tasks (0.612 and 0.583), which results in a higher positive-class F1. MiMo-V2.6-Flash is more conservative, particularly for threat classification, where it recovers only 38.8% of the contested threat sentences but with higher precision (0.561).

Qualitative inspection of the disagreement subsets did not reveal clear-cut error patterns that separate the two approaches. Both models succeed on some and fail on other sentences of the same kind, notably sentences focused on sustainability, sentences dealing with legislation, climate litigation, and activism, and sentences whose framing can only be recognized from a broader context. Identifying systematic differences between the approaches would likely require an in-depth linguistic analysis of the disagreement cases.

6 Discussion

The central finding of this study is that only the most recent prompting model closes the performance gap to fine-tuned encoders. Prior comparisons have consistently found that fine-tuned encoders outperform zero- and few-shot LLM prompting when sufficient training data is available (Bucher and Martini, 2024; Galke and Scherp, 2024; Edwards and Camacho-Collados, 2024; Zhang et al., 2025). Our results for Llama 4 Maverick, released in April 2025, and the smaller Qwen3.8-27B, released as open weights on August 14, 2026, follow this pattern: both remain below GBERT-large with ±4 context and only reach the level of GBERT-large with ±1 context. Only MiMo-V2.6-Flash, released in September 2026, matches the best encoder. Conclusions about the relative merits of fine-tuning and prompting are therefore highly time-dependent. Because release date, size, and architecture vary jointly across the evaluated models, we cannot attribute MiMo-V2.6-Flash's advantage to a single factor.

The context experiments support the argument that framing is often expressed across sentence boundaries rather than within a single sentence (Matthes and Kohring, 2008). Every extension of the context window improved all encoders, with diminishing but still positive returns, so the optimal window size beyond ±4 remains an open question.

Model size appears to matter more than the pre-training objective: the two large encoders perform similarly, while the base-sized GeistBERT consistently falls behind.

On the contested sentences, neither of the two best models is reliably more accurate, and both fail on sentences of the same kind. Many of these cases may be genuinely ambiguous, which may partly reflect the sentence-level labels inherited from longer coded segments (Section 3). Because the two approaches agree on more than 80% of sentences, their agreement could serve as a practical quality signal, for instance by routing only disagreement cases to manual coding, although the reliability of agreement cases remains to be tested.

The natural-distribution evaluation qualifies the parity observed on the balanced sets. GBERT-large loses considerable precision when confronted with the full set of neutral sentences. A plausible explanation is the shift in class priors: the encoders were trained on balanced data and thus implicitly assume a 50% base rate, whereas the prompting models are not exposed to any training distribution. Encoders applied to full corpora should therefore be trained on the natural distribution or have their decision threshold adjusted.

Jev-1.13.0 yields a mixed picture, with competitive and highly stable solution classification but the weakest threat classification. Possible explanations include its primarily English training and the absence of generated reasoning, which may matter more for threat framing, where the framing is often implicit, but our experiments cannot distinguish between them.

In practical terms, fine-tuned encoders require labeled data and GPU resources for training, but afterwards offer fast, deterministic, and locally executable inference that remains reproducible over time. Prompting requires no labeled data but depends on API access to models that may change or be retired. Notably, MiMo-V2.6-Flash was also the cheapest generative model in our comparison, with all runs costing approximately USD 2.3, compared to USD 12.4 for Llama 4 Maverick and USD 13.1 for Qwen3.8-27B.

7 Conclusion

We compared fine-tuned encoder models and zero-shot prompting for classifying threat and solution framing in sentences from German-language climate news, using five-fold cross-validation with article-grouped folds. The best encoder, GBERT-large with ± 4 context, and the best prompting model, MiMo-V2.6-Flash, performed on par, with weighted F1 scores of 0.805 and 0.802 versus 0.807 and 0.798 for threat and solution classification, respectively. Prompting models released before MiMo-V2.6-Flash remained below the best encoder, indicating that this parity is a recent development. Additional context improved all encoders consistently.

For sentence-level framing classification, current prompting models have thus become a viable, low-cost alternative to fine-tuned encoders even when labeled training data is available, while fine-tuned encoders remain attractive where reproducibility and local execution matter. Future work should examine context windows beyond four sentences, training encoders on the natural class distribution, and agreement-based prioritization of sentences for manual coding, and such comparisons should be repeated as new models are released.

Limitations

The sentence-level labels are inherited from manually coded segments, which introduces label noise that likely lowers the attainable performance for all models, and no formal intercoder reliability statistics are available for the annotations (Krippendorff, 2004). The balanced undersampling discards potentially informative negative examples and, as the natural-distribution evaluation shows, reduces the precision of the fine-tuned encoders. Learning rate and number of epochs were set per model and input configuration without a separate validation split, which may slightly favor the encoders. No formal significance tests were conducted, and several of the reported differences lie within the standard deviation across folds. For the prompting models, we evaluated a single zero-shot prompt design with one pass per model, and we cannot rule out that the models were exposed to the published news articles during pre-training, although not to the annotations. The use of API-based models, including the closed-weights Jev-1.13.0, limits long-term reproducibility (Carlson et al., 2026). Finally, the corpus is restricted to two Austrian newspapers and a six-month period, and it remains to be tested whether the findings generalize to other outlets, languages, and framing categories.

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.

Raven Adam, Marie Kogler, and Martina Scholger. 2023. Aufwind in der Berichterstattung zum Klimaschutz: Langfristige Entwicklung von Themen und Stimmungsbildern in österreichischen Zeitungen. *Zeitschrift für Digitale Geisteswissenschaften*, 8.

Muskan Nawaz Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of EMNLP 2022*.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, et al. 2025. Open-source LLMs for text annotation: A practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8:17.

Natalia Badullovich, Will J. Grant, and Rebecca M. Colvin. 2020. Framing climate change for effective communication: A systematic map. *Environmental Research Letters*, 15(12):123002.

Jelle W. Boumans and Damian Trilling. 2016. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4(1):8–23.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Manuel J. J. Bucher and Mario Martini. 2024. Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification. *arXiv preprint arXiv:2406.08660*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL-IJCNLP 2015*, pages 438–444.

David Carlson et al. 2026. The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of COLING 2020*, pages 6788–6796.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3693–3699.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451.

Fabian Dablander, Sophia Wimmer, et al. 2025. Media coverage of climate activist groups in Germany. *Climatic Change*, 178(8):144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Zackary O. Dunivin. 2025. Scaling hermeneutics: A guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, 14:28.

Alexander Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of LREC-COLING 2024*.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Mike Farjam, Hendrik Meyer, and Moritz Lohkamp. 2025. A practical guide and case study on how to instruct LLMs for automated coding during content analysis. *Social Science Computer Review*. Online-First.

Lauren Feldman and P. Sol Hart. 2018. Is there any hope? How climate change news imagery and text influence audience emotions and support for climate mitigation policies. *Risk Analysis*, 38(3):585–602.

Lauren Feldman, P. Sol Hart, and Tijana Milosevic. 2015. Polarizing news? Representations of threat and efficacy in leading US newspapers' coverage of climate change. *Public Understanding of Science*, 26(4):481–497.

Lukas Galke and Ansgar Scherp. 2024. Are we really making much progress in text classification? A comparative review. *arXiv preprint arXiv:2204.03954v6*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences (PNAS)*, 120(30):e2305016120.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Lars Guenther, Michael Brüggemann, and Sherine Elkobros. 2022. From global doom to sustainable solutions: International news magazines' multimodal framing of our future with climate change. *Journalism Studies*, 23(1):131–148.

Lars Guenther, Susan Jörges, Daniela Mahl, and Michael Brüggemann. 2024. Framing as a bridging concept for climate change communication: A systematic review based on 25 years of literature. *Communication Research*.

P. Sol Hart and Lauren Feldman. 2014. Threat without efficacy? Climate change on U.S. network news. *Science Communication*, 36(3):325–351.

Valerie Hase, Daniela Mahl, Mike S. Schäfer, and Tobias R. Keller. 2021. Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). *Global Environmental Change*, 70:102353.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of CoNLL 2019*, pages 504–514.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of COLING 2020*, pages 904–914.

David Maier. 2025. Automated detection of threat- and solution-oriented climate change framing in German-language newspapers. Master's thesis, University of Graz.

Jörg Matthes and Matthias Kohring. 2008. The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279.

Liang Mei, Jing Yao, Yutao Ge, Yan Wang, Bin Bi, Yunhai Cai, Jia Liu, Minghua Li, Zhiwei Li, Dongsheng Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*.

Hendrik Meyer, Mike Farjam, Hannah Rauxloh, and Michael Brüggemann. 2025. From disruptive protests to disrupted news frames: Comparing German news on climate protests. *Journalism*.

Itay Mondshine, Tamar Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Proceedings of LoResMT 2025 (ACL Workshop)*, pages 81–104.

Matthew C. Nisbet. 2009. Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development*, 51(2):12–23.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of SemEval-2023*, pages 2343–2361.

Qwen Team. 2026. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8. Model card: https://huggingface.co/Qwen/Qwen3.8-27B.

Sophie Elena Riesinger. 2024. Mixed-methods analysis of threat- and solution-oriented statements in Austrian newspaper articles on climate change. Master's thesis, University of Graz, Graz, Austria.

Stefan Ringhofer, Annina Thaller, Laura S. Loy, and Marie Lisa Kogler. 2026. Comparing threat- and solution-oriented climate communication: The role of hope, self-efficacy, and climate anxiety in motivating high-impact climate-friendly behavior. *Journal of Environmental Psychology*, 115:103194.

Raphael Scheible-Schmitt and Johann Frei. 2025. GeistBERT: Breathing life into German NLP. In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 42–50, Varna, Bulgaria. INCOMA Ltd.

Sander Schulhoff, Michael Ilie, Nishant Balepur, et al. 2024. The prompt report: A systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2023. A dataset for detecting real-world environmental claims. In *Findings of ACL 2023*.

Katja Thier and Tai-Tse Lin. 2022. How solutions journalism shapes support for collective climate change adaptation. *Environmental Communication*, 16(8):1027–1045.

Katja Thier and Xue Wu. 2024. Framing climate solutions: An exploratory quantitative content analysis. *Environmental Communication*, 19(3):359–375.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 5998–6008.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBERT: A pretrained language model for climate-related text. In *Proceedings of the AAAI Fall Symposium*. ArXiv:2110.12010.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New encoders for German trained from scratch: Comparing ModernGBERT with converted LLM2Vec models. *arXiv preprint arXiv:2505.13136*.

Jiacheng Zhang, Yiding Huang, Shanshan Liu, Yongqiang Gao, and Xueqi Hu. 2025. Do BERT-like bidirectional models still perform better on text classification in the era of LLMs? In *Findings of EMNLP 2025*. ArXiv:2505.18215.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Retrieving Climate Change Adaptation Concepts from World Heritage Management Plans in the Netherlands

Vedant Jumle[1], Orhan Ağaoğlu[1], Laurens Mol[1], Edgar Salas Gironés[2,3], Nan Bai[1]
[1]Delft University of Technology, Delft, the Netherlands
{vvjumle, oagaoglu1, L.P.J.Mol}@student.tudelft.nl, n.bai@tudelft.nl
[2]Technopolis Group Amsterdam, Amsterdam, the Netherlands
edgar.girones@technopolis-group.com
[3]University of Amsterdam, Amsterdam, the Netherlands
e.salasgirones@uva.nl

Abstract

Monitoring climate change adaptation (CCA) in the heritage sector, especially in the context of World Heritage (WH) management, requires systematic analyses of policy documents for CCA strategies, a process that is difficult to scale. We ask whether automated retrieval systems can *prioritise* expert annotation effort: given a CCA concept, can a system rank passages of a management plan so that those an expert would annotate appear near the top, without any in-domain supervision? This objective is formulated as an information retrieval problem with a new benchmark dataset on climate-adaptive heritage management strategies, created from expert-labelled text chunks in official Management Plans of 11 Dutch WH properties. Four retrieval approaches (BM25, a bi-encoder, a dense retriever with reranker, and a cross-encoder reranker) are evaluated and compared. The cross-encoder reranker achieves the strongest performance across labels. An LLM-as-judge evaluation further shows that top-ranked unlabelled chunks could be relevant, suggesting that the existing annotations are precise but not exhaustive. Per-label analysis identifies which concepts are automatically retrievable and which remain beyond the reach of general-purpose retrieval. Our benchmark establishes a foundation for scalable mapping and monitoring of climate change adaptation in the context of heritage management.

1 Introduction

Monitoring climate change adaptation (CCA) at scale requires structured analysis of policy documents in which adaptation strategies are recorded. In the context of UNESCO World Heritage (WH), Management Plans (MPs) are one such instrument: World Heritage managers use them to translate risks into conservation responses. Heritage properties face compounding and accelerating climate pressures (Sesana et al., 2021). Chen et al. (2025b) find empirically that 80% of UNESCO sites already experience harmful heat- and moisture-related climate risks. Climate risks and adaptation strategies are increasingly being addressed in MPs as heritage properties are especially vulnerable amidst global climate change. Properties worldwide are also expected to develop tailored climate action plans by 2030 (UNESCO, 2023; Chatziandreou and Haupt, 2026). Yet policy documents systematically fail to translate this awareness into operational actions. Analysing 1,868 World Heritage documents from 535 locations, Chen et al. (2025a) find a persistent awareness-action gap that sites with greater vulnerability recognition implement fewer concrete local management interventions. Fatorić and Biesbroek (2020) document that in the Netherlands, the largest perceived barrier for heritage sites to adapt to climate change impacts is lack of adaptation policy for heritage, and Guzman (2025) identifies similar challenges of cross-sector synergy in Norwegian national climate planning. Cheang et al. (2025a) developed a domain-expert taxonomy of 27 climate-adaptive heritage management strategy codes and manually annotated hundreds of sentences from 11 World Heritage MPs in the Netherlands, thereby providing a reusable and systematic evidence base for understanding and assessing CCA in heritage management at scale. This annotation is accurate but expert-intensive and cannot readily extend to the growing corpus of heritage policy documents worldwide, a shortfall made more acute by the scale and urgency of the underlying climate problem.

Parallel progress in Natural Language Processing (NLP) and Information Retrieval (IR) for climate policy text provides a methodological potential for addressing this challenge. Webersinke et al. (2021) show that domain-adapted language models, continued-pretrained on climate-related text, substantially outperform general-purpose models on climate text classification. Bonenkamp et al. (2025) demonstrate that fine-tuned encoder-only models achieve the strongest results on adaptation-

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 170–182
October 28, 2026 ©2026 Association for Computational Linguistics

relevance classification in climate policy text, with text granularity as a critical design factor. Schimanski et al. (2024) release ClimRetrieve as a retrieval benchmark, covering over 8,500 question-answer pairs from corporate sustainability reports, and show that generic embeddings require expert augmentation to handle domain-specific language. The scale of the challenge motivates automation. The IPCC AR6 WGII alone drew on more than 34,000 papers (De Rijke et al., 2025), and Sietsma et al. (2024) show that ML-based analysis of 8,691 adaptation policy documents surfaces geographic patterns that manual review cannot detect.

While these methods have been developed and evaluated on climate corpora, applying them to practitioner-authored heritage policy documents remains largely unexplored. Hall and Ram (2016) demonstrate through lexical analysis of all five IPCC assessment reports (1990–2014) that heritage terminology is minimally present in IPCC discourse, reflecting a conceptual separation between the two domains that extends to the vocabulary level. Chen et al. (2025a) use keyword search and LLM classification to respectively evaluate climate awareness and IPCC local action in WH properties, but the analyses on both domains are separate. Bai et al. (2025) apply sentence embeddings to identify shared CCA topics across Dutch WH properties, but do not systematically validate the obtained similarity scores. We address this cross-domain evaluation gap directly. The manually labelled sentence-level dataset of climate-adaptive heritage management strategies in Cheang et al. (2025a) is treated as the ground-truth retrieval reference. We formalise the evaluation process using comparable metrics and test four common IR approaches, with the following question:

Given a CCA concept and its expert definition, can an automated retrieval system rank the text chunks of a World Heritage MP so that those expert-annotated passages with that concept appear near the top, without any in-domain supervision?

This effectively functions as a benchmark[1] for ClimateNLP researchers to optimise retrieval methods and for heritage scholars to potentially scale up the analysis on the intersection of CCA and heritage management.

[1]Code, the extracted chunks, the concept definitions, the alignment output and the LLM-as-judge prompts and judgements are available at https://github.com/vedant-jumle/Heritage-Doc-Retrival.

2 Methodology

2.1 Datasets and Labels

Using all the textual data in MPs from 11 Dutch WH properties $\boldsymbol{\mathcal{H}} := \{\mathcal{H}_l | l = 1, 2, ..., n\}$ (see Table A2), two separate datasets were constructed:

Definition 1 (Sentence-level Labelled Dataset $\mathcal{D}_1$)**.** Let $\mathcal{D}_1 = (\mathcal{T}_1, \mathcal{C}, \mathcal{L})$, where:

- $\mathcal{T}_1 = \{t_1^{(i)(l)} | i = 1, 2, ..., M, t_1^{(i)(l)} \in \mathcal{H}_l\}$ is the set of $M = 763$ sentences manually extracted from the MP source texts.
- $\mathcal{C} = \{c^{(k)} | k = 1, 2, ..., P\}$ is the set of $P = 27$ *climate-adaptive heritage management strategy labels* assigned by experts.
- $\mathcal{L} : \mathcal{T}_1 \rightarrow \mathcal{C}$ is the manual labelling function that maps each sentence to one label.

$\mathcal{D}_1$ originates from Cheang et al. (2025b), who manually reviewed MPs from 11 Dutch WH properties and came up with an inductive coding system of 27 climate-adaptive heritage management strategies in the Netherlands (see Table A1). These 27 labels encompass both strategies and operational considerations related to CCA. The strategies are categorised into past, present, and future approaches, including '*Adaptation*', '*Conservation*', and '*Knowledge Transfer*'. The operational considerations comprise challenges, such as '*Climate Change*' and '*Soil Subsidence*', as well as managerial considerations, such as '*Management Variation*' and '*Environmental Variation*'.

Definition 2 (Chunk-level Unlabelled Dataset $\mathcal{D}_2$)**.** Let $\mathcal{D}_2 = (\mathcal{T}_2, \mathcal{C}_{\text{def}})$, where:

- $\mathcal{T}_2 = \{t_2^{(j)(l)} | j = 1, 2, ..., N, t_2^{(j)(l)} \in \mathcal{H}_l\}$ is the set of N text chunks extracted from all MPs and indexed by a unique Chunk ID j.
- $\mathcal{C}_{\text{def}} = \{[c^{(k)}, d^{(k)}] | k = 1, 2, ..., P'\}$ is a smaller set of $P' = 24$ labels, where $d^{(k)}$ is an expert-authored definition for concept $c^{(k)}$.

$\mathcal{D}_2$ originates effectively from the same sources of MPs as $\mathcal{D}_1$, but with a different format to organize information. The MPs are digitally authored PDFs with no scanned pages. We extract text $\mathcal{T}_2$ using Docling (Auer et al., 2024) with GPU-accelerated layout analysis while OCR (Optical Character Recognition) is disabled, producing clean markdown output free of the character-level artefacts (e.g., vertical logo text, page number bleed-through) common in naïve PDF parsing. Each extracted document is split into text chunks

using a sliding window over characters with a default setting of `chunk_size`$= 1000$ and an overlap of `chunk_overlap`$= 200$, snapping boundaries to the nearest sentence end to avoid mid-sentence splits. This yields $N = 2{,}926$ chunks across 10 MPs ($\mathcal{H}_9$ from *Beemster*[2] is excluded throughout, since this is the only Dutch-language MP). Chunk size is a hyperparameter with non-trivial effects on retrieval, as larger chunks reduce the candidate pool and provide more context per match, while smaller chunks offer finer granularity. We report a chunk-size ablation over six log-spaced configurations (200–3,000 characters) in Section 3. Three labels (*Environmental variation*, *Preservation*, and *Restricted greenery*) are excluded from $\mathcal{C}_{\text{def}}$ since they lack clear expert-authored definitions or representativeness in our dataset (see Table A1).

Because $\mathcal{T}_2$ was drawn from the same source documents as $\mathcal{T}_1$, every sentence $t_1^{(i)(l)} \in \mathcal{T}_1$ is fully or partially contained within one chunk $t_2^{(j)(l)} \in \mathcal{T}_2$ or two near-duplicate chunks from the same MP $\mathcal{H}_l$. The two datasets are linked at the text level through an alignment function ψ described below.

Definition 3 (Text Alignment ψ)**.** A *text alignment* maps each labelled sentence in $\mathcal{T}_1$ to the set of one or more chunks in $\mathcal{T}_2$ that contain it.

For each labelled sentence $t_1^{(i)(l)}$, we locate the chunk(s) in $\mathcal{T}_2$ that contain(s) it using a five-pass matching procedure: (i) exact substring match, (ii) split-pair match (first or second half of the sentence), (iii) split-triplet match, (iv) fuzzy match (token-sort ratio ≥ 85), and (v) fuzzy split match. Sentences with no match under any pass are recorded as unmatched. Every labelled sentence from the 10 evaluated MPs is matched, while 32 unmatched sentences all belong to *Beemster* ($\mathcal{H}_9$), which is excluded throughout. In our experiment, the alignment is dominated by exact substring matches (95.3% of matched sentences). The remaining 4.7% rely on the split and fuzzy passes, which are the only passes that can in principle produce a false-positive alignment. Exact matches cannot misalign by construction, so alignment noise is limited to the 4.7% of matches produced by the split and fuzzy procedures, and we release these 35 approximate matches alongside the code for inspection. Since the labelled dataset $\mathcal{D}_1$ has been reviewed and annotated using the same corpora $\boldsymbol{\mathcal{H}}$ as the retrieval dataset $\mathcal{D}_2$, there could naturally be an implicit assumption suggesting that those blocks not associated with any labelled sentences would be negative examples, not relevant to any concepts in $\mathcal{C}_{\text{def}}$, which we will challenge with our experiments. Using the mapping function ψ, the labelled dataset $\mathcal{D}_1$ can be aligned with $\mathcal{D}_2$, such that:

Definition 4 (Aligned Chunk-level Labelled Dataset $\mathcal{D}_2'$)**.** Let $\mathcal{D}_2' = (\mathcal{T}_2, \mathcal{C}_{\text{def}}, \mathcal{L}')$, where:

- $\mathcal{T}_2$ is the same set of $N = 2{,}926$ text chunks.
- $\mathcal{C}_{\text{def}}$ is the same set of $P' = 24$ concepts enhanced with their definitions.
- $\mathcal{L}' : \mathcal{T}_2 \to 2^{\mathcal{C}_{\text{def}}}$ is the function that maps each text block to none, one or more labels.

Formally, $\mathcal{D}_2'$ is potentially multi-label (71 chunks, 2.4%), since one chunk could contain sentences related to different labels, meaning that a labelled chunk can result in several valid label-chunk pairs.

2.2 Problem Definition

We address the following problem of retrieval consistency: given a heritage strategy label $c^{(k)} \in \mathcal{C}_{\text{def}}$ and its definition $d^{(k)}$, our goal is to identify the subset of text chunks in each MP that are relevant to that label. This requires a method to assign a relevance score to every chunk, so that chunks can be ranked by their relevance to a given concept. We evaluate how well such rankings agree with expert annotation. Converting these scores into a variable-size relevant set per concept-MP pair is left to future work. We formalise this as a retrieval function f for the aligned dataset $\mathcal{D}_2'$, where: $f : \mathcal{C}_{\text{def}} \times \mathcal{T}_2 \to \mathbb{R}$ assigns a score to every label-chunk pair. Figure 1 gives an overview of the full pipeline described in this section.

2.2.1 Query Formation

Each climate-adaptive heritage management strategy label $c^{(k)}$ is represented as a single query string by concatenating the label name and its definition, formally as $q^{(k)} =$ "$c^{(k)}$: $d^{(k)}$". This gives the retrieval system access to both the label name (useful for lexical matching) and its semantic definition (useful for embedding-based matching), without requiring any fine-tuning or query expansion.

2.2.2 Ground Truth

The ground truth for retrieval systems is derived from the aligned dataset $\mathcal{D}_2'$. Multiple labelled sentences may align to the same chunk under

[2]See Table A2. *Droogmakerij de Beemster (Beemster Polder)* as a property inscribed in 1999, available at https://whc.unesco.org/document/218353, accessed 20 September 2026.

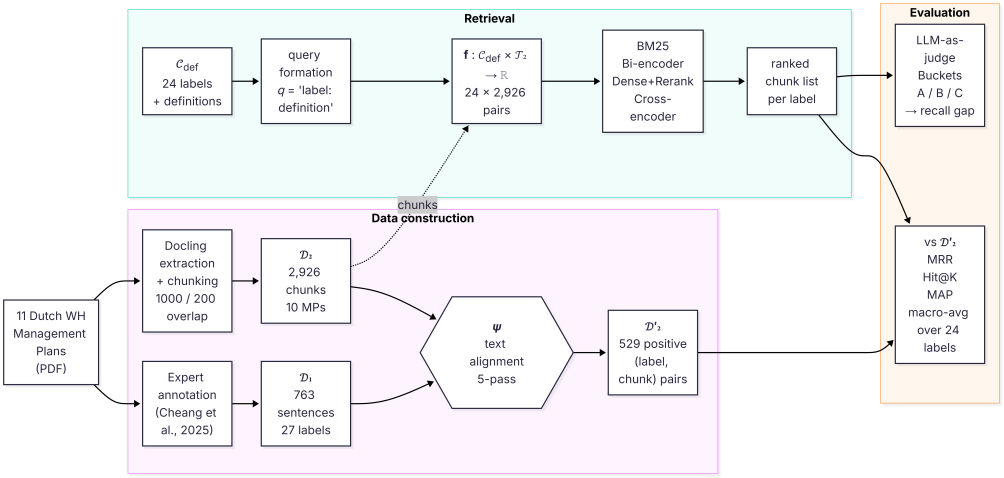


Figure 1: Overview of the pipeline. The same 11 Management Plans give rise to two datasets: the expert-annotated sentences $\mathcal{D}_1$ and the Docling-extracted chunks $\mathcal{D}_2$. The alignment ψ links them into the chunk-level ground truth $\mathcal{D}'_2$. Each concept in $\mathcal{C}_{\text{def}}$ is turned into a query, the retrieval function f scores every label-chunk pair, and the resulting ranking is evaluated both against $\mathcal{D}'_2$ and by an LLM judge.

the same label. After deduplication, 529 unique query-chunk pairs remain across 24 labels. A detailed explanation of how the original 763 labelled sentences become 529 query-chunk pairs is provided in Appendix A. The retrieval function f is evaluated against the human-annotated labels in $\mathcal{D}'_2$, measuring how consistently automated ranking agrees with expert judgement. For each label $c^{(k)} \in \mathcal{C}_{\text{def}}$, the retrieval function f scores all $N = 2{,}926$ chunks in $\mathcal{T}_2$ and produces a ranked list of chunks ordered by their retrieval scores. The ground truth for label $c^{(k)}$ is the set of positive chunks $\mathcal{G}^{(k)} = \{t_2^{(j)(l)} \mid \mathcal{L}'(t_2^{(j)(l)}) \ni c^{(k)}\}$ derived from the alignment ψ. Evaluation metrics are then computed over this ranked list against $\mathcal{G}^{(k)}$, and macro-averaged across all P' evaluable labels.

2.2.3 Retrieval Systems

Four common retrieval systems of increasing sophistication are evaluated, spanning the lexical–semantic spectrum.

BM25. A bag-of-words lexical baseline using the *Okapi BM25* algorithm (Robertson and Zaragoza, 2009). Chunks are tokenised by whitespace. BM25 scores are computed for each query–chunk pair and used to rank all N chunks. BM25 requires no training and serves as a strong lexical baseline that directly rewards surface-form overlap between query terms and chunk text.

Bi-encoder. A dense retrieval system based on `all-mpnet-base-v2` (Song et al., 2020), a general-purpose sentence encoder pre-trained on over one billion sentence pairs. Query and chunk texts are encoded independently into 768-dimensional vectors and ℓ_2-normalised, while scores are computed as cosine similarities. The bi-encoder captures semantic similarity beyond surface overlap, at the cost of treating query and chunk in isolation.

Dense + Rerank. A two-stage pipeline: the bi-encoder first retrieves the top-50 chunks by cosine similarity, which are then reranked by a cross-encoder `ms-marco-MiniLM-L-6-v2` (Reimers and Gurevych, 2019). The cross-encoder scores each query-chunk pair jointly, enabling fine-grained relevance judgement at the cost of higher inference time. Only the top-50 candidates are passed to the cross-encoder to keep inference tractable.

Cross-encoder. The cross-encoder is also evaluated as a standalone system, scoring all $P' \times N = 24 \times 2{,}926 = 70{,}224$ query-chunk pairs jointly without any bi-encoder pre-filter. This approach is well-suited to our goal, as it assigns a fine-grained relevance score to every chunk in the corpus for a given query. Since it produces a comparable score for every pair rather than only an ordering within a shortlist, it is also the system best suited to a future thresholding step. The trade-off is computational cost, which remains feasible at the scale of this corpus, but would suffer when further scaled up.

2.3 Evaluation

2.3.1 Evaluation Metrics

We evaluate with rank-based metrics, which measure if the systems assign high scores to known-

relevant chunks so that they surface near the top of the ranking. To quantify the evaluation, we adopt three standard ranking metrics in the IR literature.

First, *mean reciprocal rank* (MRR) captures the rank of the first positive chunk for each label, averaged across all evaluated labels:

$$\text{MRR} = \frac{1}{P'} \sum_{k=1}^{P'} \frac{1}{\text{rank}_k^*},$$

in which rank_k^* is the rank of the highest-scoring positive chunk for label $c^{(k)}$ in the retrieval output. MRR penalises systems that rank a known-relevant chunk lower, directly reflecting the goal of surfacing relevant content near the top. We complement MRR with Hit@K, which measures if at least one positive chunk appears within the top-K results:

$$\text{Hit@}K = \frac{1}{P'} \sum_{k=1}^{P'} 1[\text{rank}_k^* \leq K],$$

where $1[\cdot]$ is the indicator function and we report $K \in \{1, 3\}$, while a larger range of K is left for future studies. Hit@K is used since multiple chunks may be positive for a single label (due to conceptual overlap across passages), and surfacing any one of them within the top K is practically useful.

Lastly, *mean average precision* (MAP) rewards systems that rank *all* known positive chunks highly:

$$\text{MAP} = \frac{1}{P'} \sum_{k=1}^{P'} \text{AP}_k,$$

$$\text{AP}_k = \frac{1}{|\mathcal{G}^{(k)}|} \sum_{r=1}^{N} P_k(r)\, \text{rel}_k(r),$$

where $\mathcal{G}^{(k)}$ is the set of positive chunks for label $c^{(k)}$, N is the total number of ranked chunks, $P_k(r)$ is the precision at rank r, and $\text{rel}_k(r) = 1$ if the chunk at rank r is in $\mathcal{G}^{(k)}$ and 0 otherwise. MAP is particularly suitable when multiple chunks per label are relevant, as it measures retrieval completeness rather than just the top hit.

2.3.2 LLM-as-Judge Evaluation

The original dataset $\mathcal{D}_1$ may contain several expert-labelled sentences per label per MP. However, because the labels were provided through manual read-through, experts may have unintentionally missed other relevant sentences. This means that the ground truth has high precision but inherently low recall, where genuinely relevant chunks that were not selected during annotation are treated as negatives. To quantify this recall gap, we employ an LLM-as-judge protocol (Zheng et al., 2023). For each of the 24 labels, we sample three buckets of label-chunk pairs: Bucket A (ground-truth positive chunks), Bucket B (top-ranked chunks *not* in the ground truth), and Bucket C (low-ranked chunks not in the ground truth). Each pair is scored 1–5 for relevance by an LLM[3] using a structured prompt that provides the label name, its definition, and the chunk text (see Appendix A.3). The intended design is 2 from Bucket A, 3 from Bucket B and 1 from Bucket C per label, giving 144 pairs in total. Because of a shortage of samples for two labels as described in Appendix A.3, we obtained 140 pairs (47 Bucket A, 70 Bucket B and 23 Bucket C pairs), scored using 50 concurrent API calls. Bucket B chunks with scores of at least 4 indicate those that the system ranked highly that ground truth does not credit, thus potentially false-negative candidates.

3 Results

3.1 Overall Retrieval Performance

Table 1 reports macro-averaged metrics across the $P' = 24$ evaluated labels for each retrieval system. The Cross-encoder obtains the highest value on every metric by ranking all $N = 2{,}926$ chunks jointly against each query without any pre-filtering. With only 24 labels, however, the margins are not all equally reliable. Bootstrapping over labels (10,000 replicates) further separates the Cross-encoder from BM25 and the Bi-encoder on MRR ($+0.13$, 95% CI $[+0.03, +0.26]$ and $+0.22$, CI $[+0.07, +0.39]$), but the Hit@1 and MAP differences are indistinguishable from zero for every pair of systems, as are all differences between the Cross-encoder and Dense + Rerank. We therefore rest the comparison on MRR and Hit@3 and do not read the small MAP differences as evidence of ranking quality. Dense + Rerank scores lower than the Cross-encoder alone, indicating that the Bi-encoder pre-filter discards relevant chunks before the Cross-encoder can evaluate them, while BM25 outperforms the Bi-encoder on most metrics. Surface-form overlap between label definitions and chunk text is a strong signal here, and independent query and chunk encoding with a generic embed-

[3] `google/gemma-4-31B-it-qat-w4a16-ct`, a 31B instruction-tuned model quantised to 4-bit weights, served on-premises through the TULIP API (an OpenAI-compatible Chat endpoint) at temperature 0. Running the judge locally avoids sending heritage policy text to third-party providers.

ding model is shown as insufficient for this inter-domain cross-vocabulary task.

System	MRR	Hit@1	Hit@3	MAP
BM25	0.32	0.21	0.33	0.13
Bi-encoder	0.23	0.17	0.21	0.13
Dense + Rerank	0.36	0.21	0.50	0.12
Cross-encoder	**0.45**	0.29	**0.54**	0.14

Table 1: Macro-averaged retrieval metrics across 24 predefined labels. Bold marks the Cross-encoder's MRR and Hit@3 as the only margins a label-level bootstrap separates from BM25 and the Bi-encoder. No metric separates the Cross-encoder from Dense + Rerank, and Hit@1 and MAP do not separate any pair of systems.

3.2 Per-Label and Per-Document Analysis

Performance varies substantially across labels (Table 2). Seven climate-adaptive heritage management strategy labels achieve Cross-encoder MRR of 1 with varying MAP scores, indicating that for these concepts the annotated chunks consistently rank highly. For *Maritime management*, the Cross-encoder misses Hit@1 (with MRR of 0.50, i.e. the first positive chunk appears at rank 2), but its sufficiently high MAP of 0.50 indicates that both of its positive chunks are still ranked highly. At the other extreme, five labels score Cross-encoder MRR lower than 0.06, suggesting either ambiguous concept definitions or insufficient label-specific content in the MPs. Combined with Table A1, it is evident that the frequency of concepts is not the main constraint for retrieval systems to perform.

Notably, system rankings sometimes invert for individual labels. BM25 achieves MRR of 1 on *Restoration*, where the ground-truth chunk contains the keyword directly. Dense + Rerank, however, scores 0.02, and the Cross-encoder recovers only to 0.50 by de-prioritising it in favour of semantically adjacent but lexically distinct chunks. Conversely, *Maritime management* yields Bi-encoder MRR of 1 against BM25's 0.10, a case where semantic similarity succeeds and keyword overlap fails.

Performance also varies substantially across the 10 evaluated MPs (see Table A2 for basic statistics and Table A3 for detailed MAP scores per metric), as shown in Figure 2. *Schokland* ($\mathcal{H}_7$) is the easiest property for the three systems other than the Cross-encoder (BM25 MAP of 0.39, Bi-encoder and Dense + Rerank MAP of 0.37). It is the smallest MP (28 Docling chunks) and its content closely matches the vocabulary of the relevant strategy labels. *Rietveld Schröder House* ($\mathcal{H}_{10}$) achieves the highest Cross-encoder MAP (0.35) despite moderate BM25 performance (0.25), suggesting its content is semantically well-aligned with label definitions but lexically distinct. *Van Nellefabriek* ($\mathcal{H}_6$) is the most challenging property for BM25 and the Cross-encoder (MAP of 0.10 and 0.09 respectively), as it covers only 4 evaluable labels and its industrial heritage vocabulary diverges from the climate-adaptive management strategy definitions. The Bi-encoder scores lowest on *Colonies* (0.06) and Dense + Rerank on *Amsterdam* (0.07). *Amsterdam Canal Ring* ($\mathcal{H}_5$) and *Dutch Water Defence Lines* ($\mathcal{H}_3$) show below-average performance across all systems, indicating that large, topic-diverse MPs dilute the label-specific signal.

Label	MRR	Hit@3	MAP
Top 7 (Cross-encoder)			
Soil subsidence	1.00	1	0.54
Limited heritage inclusion	1.00	1	0.26
Climate change	1.00	1	0.25
Management variation	1.00	1	0.22
Human activities	1.00	1	0.17
Fisheries management	1.00	1	0.17
Stakeholder collaboration	1.00	1	0.07
Maritime management	0.50	1	0.50
Bottom 5 (Cross-encoder)			
Legislation	0.06	0	0.05
Research	0.05	0	0.02
Prohibition	0.01	0	0.00
Coherence	0.01	0	0.01
Research gap	0.01	0	0.01

Table 2: Best and worst performing labels under the generally best-performing Cross-encoder system.

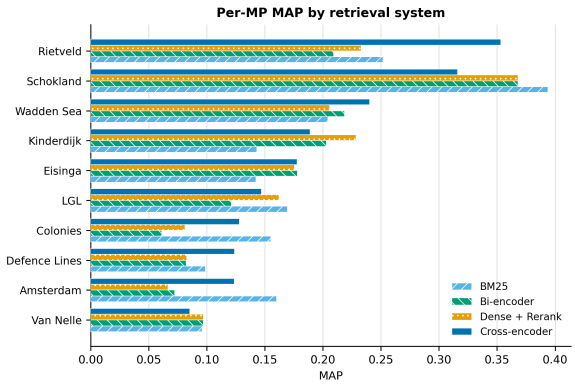


Figure 2: Per-MP MAP for all systems. MPs are sorted by the scores on Cross-encoder MAP.

3.3 Ablation Study on Chunk Sizes

We ablate chunk size while holding all other parameters fixed. As shown in Figure 3, all four systems

are evaluated at each configuration. BM25 and Dense + Rerank trend upward with chunk size. The Cross-encoder peaks at chunk size of 1,384 (MRR of 0.54) and then declines slightly at 3,000, while the bi-encoder improves gradually. At chunk size of 3,000, BM25 reaches MRR of 0.46 and Dense + Rerank reaches 0.42, substantially above the default configuration. However, larger chunks mechanically reduce the candidate pool (904 chunks at 3,000 vs. 2,881 at 940) and pack more text per chunk, making it easier to contain ground-truth sentences, conflating retrieval quality with corpus granularity. This confound also affects any *relative* quantity computed over chunks, including alignment match rates, since larger chunks absorb more labelled sentences by construction. Such ratios are thus not comparable across chunk sizes. We treat the ablation as a sensitivity analysis over independently re-chunked corpora, and retain the 1,000-character configuration for all primary results.

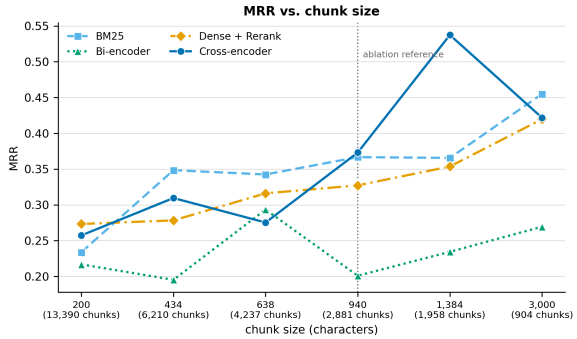


Figure 3: MRR computed over changing chunk sizes for all systems. The dotted line marks the 940-character grid point, the ablation's closest configuration to the 1,000-character setting used for all primary results. Chunk counts are given under each grid point, since chunk size and corpus granularity vary together.

3.4 Relevance of Unlabelled Chunks

The Cross-encoder assigns a real-valued score to every query-chunk pair. Figure 4 shows the distribution of these scores separately for ground-truth positive chunks and all other chunks, aggregated across all labels. Positive chunks cluster at higher scores while negatives span the full range, but the distributions overlap substantially, reflecting the recall gap established in Section 2.3.2. Many high-scoring, unlabelled chunks can be genuinely relevant. The weak but significant Spearman correlation ($r = -0.259$, $p = 0.030$) between Cross-encoder rank and LLM relevance score for Bucket B chunks indicates that score order carries some signal in the unlabelled region while being far from a restatement of the retrieval score itself.

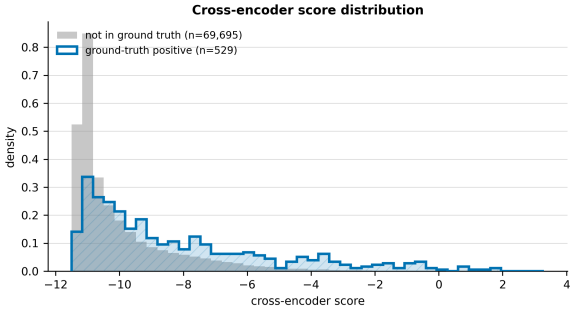


Figure 4: Cross-encoder score distribution for ground-truth positive chunks (GT) and all other chunks, aggregated across 24 labels. Distributions overlap because genuinely relevant but unannotated chunks score similarly to annotated positives.

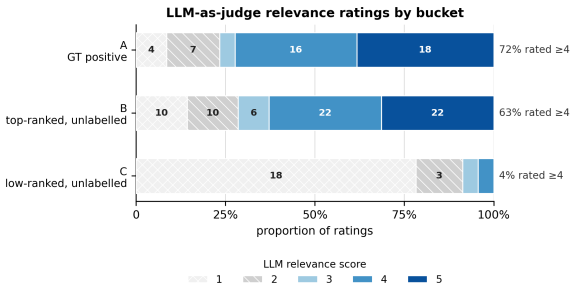


Figure 5: LLM relevance ratings by sampling bucket, shown as the proportion of ratings at each of the five discrete score levels (counts printed inside the bars, $n = 140$). Bucket B (top-ranked, unlabelled) is rated at least 4 almost as often as Bucket A (ground-truth positives), 62.9% vs. 72.3%, while Bucket C (low-ranked) is rated at least 4 only 4.3% of the time.

Figure 5 shows the full distribution of relevance scores by bucket, while the statistical effects are recorded in Table A4. High scores for Bucket A (ground-truth positives, $n = 47$) confirm that the labelled chunks are relevant, while significantly lower scores for Bucket C (low-ranked, unlabelled, $n = 23$) confirm that low-ranked chunks are genuinely irrelevant. Crucially, the scores for Bucket B (top-ranked, unlabelled, $n = 70$) are nearly as high as Bucket A, with substantial overlap in distribution and no statistically significant difference. Notably, 62.9% of top-ranked unlabelled chunks score at least 4 (relevant or highly relevant) by LLM judgment. These results demonstrate that a retrieval strategy based on a fixed-K cutoff (e.g., Hit@K) would fail to identify a large amount of relevant

content that the system correctly surfaces. The high relevance scores of top-ranked, unlabelled chunks (Bucket B) suggest that the Cross-encoder provides a strong signal for relevance that extends beyond the items in the original ground truth. This supports a future thresholding-based approach to identify the complete set of relevant chunks per document, as the gap between reported rank-based metrics (Cross-encoder MRR of 0.45) and true retrieval quality is structural, arising from the sparse labelling in $\mathcal{D}_1$, not from system failure. Table A5 in the Appendix further demonstrates a few qualitative examples of sentences from all three buckets.

4 Discussion

4.1 Reflection on the Retrieval Systems

Label names for climate-adaptive heritage management strategies are domain-specific but often literal, e.g., *Soil subsidence*, *Fisheries management*, *Stakeholder collaboration*. These terms and their definitions appear verbatim in the MPs, making surface-form overlap a genuine retrieval signal rather than a trivial one. BM25's overall MRR (0.32) exceeds the Bi-encoder (0.23), demonstrating that the vocabulary gap between heritage practice and climate science is partially bridgeable without complex semantic reasoning. This finding aligns with the observation in Schimanski et al. (2024) that lexical baselines remain competitive in specialised climate domains where terminology is consistent.

The Bi-encoder treats the query and each chunk independently, and then ranks them by cosine similarity. For broad labels such as *Coherence* or *Conservation*, the embeddings of the query would capture general semantic neighbourhoods that are not discriminative enough to separate on-topic MP passages from the background information. Without in-domain fine-tuning, the embedding space conflates semantically adjacent but task-irrelevant content. The Bi-encoder's independent encoding also misses the interaction between a label definition and specific chunk content, which Cross-encoder handles by scoring query-chunk pairs jointly.

Dense + Rerank passes only the top-50 Bi-encoder candidates, so a positive outside that window can never be scored. This is decisive for the six concepts whose positives never survive the pre-filter (Table A1). Elsewhere, the pre-filter can help by discarding distractors, and Dense + Rerank matches or exceeds the Cross-encoder on 18 of the 24 labels: the macro-averaged difference (+0.09 MRR) is driven by four labels and is not distinguishable from zero by the label-level bootstrap. The trade-off for using Cross-encoder is inference cost: with a fixed set of P' concept queries, exhaustive scoring is $O(P'N)$, i.e. linear in the number of chunks but with a large constant, since every chunk is passed through the cross-encoder for every query. At a small $N = 2{,}926$, full joint scoring is still feasible, but larger collections would require pre-filtering, which re-introduces the candidate truncation problem identified here. Increasing the number of pre-filter candidates may help balance the two, and could be further ablated.

4.2 Failure Cases based on Problem Definition

The worst-performing labels split into two distinct failure modes that require different remedies. *Research gap*, *Prohibition*, and *Maritime Management* are concepts that are rare or absent in the MPs, suggesting **corpus sparsity**. Improvement requires either more documents or augmentation with synthetic samples. Meanwhile, *Coherence* is a broad term with **concept ambiguity** that appears throughout the MPs in many contexts, making it difficult to distinguish relevant from irrelevant chunks. Improvement requires more precise query specification, definition enrichment, or query expansion, thus not necessarily a better retrieval model. Furthermore, the **recall gap** mentioned in Section 2.3.2 is a structural constraint in our problem definition. The LLM-as-judge analysis reveals that 62.9% of top-ranked, unlabelled chunks are independently judged relevant. This finding demonstrates that the Cross-encoder's relevance scores are meaningful and can be used to identify a set of relevant chunks much larger than the original ground truth. Relying on rank-based metrics like MRR or a fixed top-K retrieval would arbitrarily exclude this content. The ground truth $\mathcal{D}_1$ was constructed as a labelling instrument rather than an exhaustive relevance annotation. Therefore, instead of viewing this as a limitation for retrieval evaluation, we see it as motivation for our future task formulation: using the system's scores to determine for each concept-MP pair how many relevant chunks exist without forcing a uniform top-K retrieval. The high quality of the top-ranked unlabelled chunks suggests that a thresholding approach on the cross-encoder scores could be a viable method for achieving this (Zheng et al., 2023). Iterative human-in-the-loop reviews and updates on the labelled set of ground truth can benefit further optimisation of retrieval systems.

4.3 Limitations and Future Directions

The main objective of this report was to build a pipeline that can correctly analyse at scale World Heritage MPs and rank relevant chunks to a set of pre-defined climate-adaptive heritage management labels. Hall and Ram (2016) have emphasised the importance of identifying more effective ways to communicate the value of heritage in relation to climate adaptation, arguing that clearer articulation of heritage value would increase its prominence in climate change documentation such as IPCC reports. The findings of this research represent an initial step towards this ultimate objective.

Several limitations bound the generalisability of these results. The ground-truth labels have high precision but inherently low recall by design, and only $P' = 24$ labels are currently evaluated. Yet even the original $P = 27$ labels from Cheang et al. (2025a) would also only be a subset of the 915-concept full taxonomy provided by IPCC. Connecting the two jargon-compact domains of heritage management and CCA remain a challenge. In the current paper, only a small corpus (10 MPs, 2,926 chunks) in the Netherlands is evaluated. All queries are built with English label definitions applied to MPs written in English, ignoring the Dutch-language MP from *Beemster*. Cross-lingual mismatch may suppress recall for passages that are more fluently expressed in Dutch. Results may also not generalise well to other heritage corpora, other languages, or larger document collections of MPs for WH properties across the globe. Furthermore, all four retrieval systems operate in a zero-shot setting. Fine-tuning on in-domain label-chunk pairs may improve all metrics and may change the relative ordering of systems. No inter-annotator agreement nor cross-system consistency for chunk-level relevance was evaluated.

The retrieval quality could be improved by taking actions in a few directions. The Bi-encoder could be fine-tuned with in-domain knowledge on (label definition, positive chunk) pairs derived from the aligned dataset $\mathcal{D}_2'$, using contrastive losses with hard negatives sampled from Bucket B pairs that the LLM judges as irrelevant. This would address the vocabulary gap between label definitions and MP text that limits zero-shot embedding performance. Replacing the general-purpose encoder models in this study with a climate-domain encoder, such as ClimateBERT (Webersinke et al., 2021) continued-pretrained on climate and heritage policy text, may improve semantic alignment for retrieval at the nexus of heritage management and CCA. Label definitions could be augmented with synonyms, related terms, or example sentences drawn from $\mathcal{D}_1$, providing richer query signal particularly for labels with few matched sentences and reducing chances for false positives. As stated earlier, the retrieval system should optimally incorporate an adaptive threshold for each label-MP pair. However, operationalising a threshold requires deciding which score separates the relevant from the irrelevant. Practical approaches can include: (i) a global percentile cutoff (e.g., top-5% of scores per label), which adapts to per-label score range differences; (ii) a fixed score threshold calibrated against the distribution of known-positive scores; or (iii) using LLM-as-judge scores to calibrate the threshold on a small sample, then applying it to the full corpus. Alternatively, (ii) and (iii) can be combined by selecting a threshold that maximises agreement with LLM judge scores on a calibration sample, then applying it corpus-wide to produce a variable-size relevant set per label-MP pair rather than a fixed top-K. Once a reliable score threshold is calibrated using LLM-as-judge feedback, the pipeline can be applied beyond the 10 Dutch MPs analysed here to the broader corpus of WH management plans worldwide, enabling scalable monitoring of CCA uptake in heritage governance.

5 Conclusions

Being at the nexus of heritage management and climate governance, we formulate an information retrieval problem with a benchmark dataset to map climate change adaptation concepts in World Heritage management documents. Four retrieval systems are evaluated against expert-annotated labels, offering insights on where to further improve both the retrieval methods and the ground-truth annotation. With a ranking approach, human experts can be potentially aided to efficiently prioritise what to annotate. The study prepares for future studies in retrieving and evaluating climate adaptation strategies developed in heritage sector at scale.

Acknowledgments

This study is supported by the Climate Action Seed Fund at TU Delft. We thank the TU Delft TULIP team for providing access to the on-premises LLM infrastructure used in the evaluation. We thank the anonymous reviewers for their useful feedback.

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikola Stanisljevic, Kateryna Abramova, and Peter Staar. 2024. Docling technical report. *arXiv preprint arXiv:2408.09869*.

Nan Bai, Kai Ut Cheang, and Ana Pereira Roders. 2025. Bridging world heritage management with climate change adaptation in the netherlands using artificial intelligence. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:79–86.

Jetske Bonenkamp, Robbert Biesbroek, and Ioannis N. Athanasiadis. 2025. Transforming adaptation tracking: benchmarking transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 266–277, Vienna, Austria. Association for Computational Linguistics.

Konstantinos Chatziandreou and Wolfgang Haupt. 2026. Integrating cultural heritage and climate policy? insights from unesco world heritage cities. *European Planning Studies*, pages 1–23.

Kai Cheang, Nan Bai, and Ana Pereira Roders. 2025a. Unveiling climate-adaptive world heritage management strategies: The netherlands as a case study. *Sustainability*, 17(12):5555.

Kai Cheang, Nan Bai, and Ana Pereira Roders. 2025b. Data from: 'unveiling climate-adaptive world heritage management strategies: the netherlands as a case study'.

Yang Chen, Dayang Wang, Luchen Zhang, Chongxiao Wang, Cheng Sun, and Qi Dong. 2025a. World heritage documents reveal persistent gaps between climate awareness and local action. *Nature Climate Change*, pages 1–7.

Zihua Chen, Qian Gao, Yajing Wu, Jiaxin Li, Xiaowei Li, Xiao Li, Zhenbo Wang, and Haiyang Cui. 2025b. World cultural heritage sites are under climate stress and no emissions mitigation pathways can uniformly protect them. *Communications Earth & Environment*, 6(1):628.

Maarten De Rijke, Bart Van Den Hurk, Flora Salim, Alaa Al Khourdajie, Nan Bai, Renato Calzone, Declan Curran, Getnet Demil, Lesley Frew, Noah Gießing, and 1 others. 2025. Report on the 1st workshop on information retrieval for climate impact (manila24) at sigir 2024. In *ACM SIGIR Forum*, volume 59, pages 1–23. ACM New York, NY, USA.

Sandra Fatorić and Robbert Biesbroek. 2020. Adapting cultural heritage to climate change impacts in the netherlands: barriers, interdependencies, and strategies for overcoming them. *Climatic Change*, 162(2):301–320.

Paloma Guzman. 2025. Cultural heritage in climate planning: An analysis of the norwegian national climate documents and guidelines. *Journal of Cultural Heritage*, 74:35–47.

C Michael Hall and Yael Ram. 2016. Heritage in the intergovernmental panel on climate change assessment reports: a lexical assessment. *Journal of Heritage Tourism*, 11(1):96–104.

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*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval*, volume 3, pages 333–389. Now Publishers Inc.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524.

Elena Sesana, Alexandre S Gagnon, Chiara Ciantelli, JoAnn Cassar, and John J Hughes. 2021. Climate change impacts on cultural heritage: A literature review. *Wiley Interdisciplinary Reviews: Climate Change*, 12(4):e710.

Anne J Sietsma, Emily Theokritoff, Robbert Biesbroek, Iván Villaverde Canosa, Adelle Thomas, Max Callaghan, Jan C Minx, and James D Ford. 2024. Machine learning evidence map reveals global differences in adaptation action. *One Earth*, 7(2):280–292.

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.

UNESCO. 2023. Policy document on climate action for world heritage. Technical report, UNESCO.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Appendix

A.1 Pre-processing of Datasets

Table A1 shows the number of labelled sentences per label in $\mathcal{D}_1$, excluding the MP of *Beemster* ($\mathcal{H}_9$). Table A2 gives an overview of basic statistics for all the MPs from the Netherlands being used as the dataset in this paper.

Concept	Count	#Pairs	R@50
Adaptation	16	14	0.36
Assessment	26 (27)	22	0.14
Climate change	65 (66)	42	0.33
Coherence	28 (29)	25	0.00
Conservation	63 (64)	53	0.00
*Environmental variation**	5	—	—
Fisheries management	10	7	1.00
Framework	55 (56)	44	0.14
Greenery	19	8	0.25
Heritage inclusion	33 (34)	27	0.11
Human activities	28	12	0.17
Knowledge transfer	96 (97)	63	0.06
Legislation	52 (54)	35	0.17
Limited heritage inclusion	4	4	0.00
Management variation	5	5	0.00
Maritime management	2	2	1.00
Monitoring	45 (47)	32	0.06
Multi-level management	19	18	0.28
*Preservation**	1	—	—
Prohibition	6	6	0.00
Research	32 (33)	23	0.13
Research gap	1	1	0.00
Restoration	5	5	0.20
*Restricted greenery**	1	—	—
Soil subsidence	4	2	1.00
Stakeholder collaboration	52 (53)	39	0.13
Water management	58	40	0.17
Total / macro-avg.	731 (744)	529	0.24

*Excluded from evaluation

Table A1: Labelled sentences per concept in $\mathcal{D}_1$ (excl. Beemster), the resulting positive label-chunk pairs, and the share of those positives that survive the Bi-encoder's top-50 pre-filter. Count gives unique annotated sentences, with the number of sentence-chunk rows in parentheses where a sentence spans two overlapping chunks. #Pairs is the number of unique positive pairs after deduplication, which is the quantity retrieval is evaluated against. R@50 is recall of those positives within the Bi-encoder's top-50 candidates. The definition used for each concept to formulate the queries can be found in our GitHub repository.

To clarify the steps taken from 763 labelled sentences to 529 query-chunk pairs, we describe the procedure as follows. The source file derived from Cheang et al. (2025b) records corresponds to 763 distinct annotated sentences, of which 731 are matched onto the Docling chunks (Table A1) and 724 remain after excluding the three labels without expert definitions. These 724 annotations across $P' = 24$ labels that have at least one matched sentence yield 737 sentence-chunk rows, because 13 sentences each span two overlapping chunks. Deduplicating by (label, chunk) then gives the 529 unique positive pairs used as "ground truth".

A.2 Extended Results

The R@50 reported in Table A1 shows what share of the known positives for each concept appears within the Bi-encoder's top-50 candidates, i.e. the fraction that the Dense + Rerank pre-filter passes on to the cross-encoder. For six concepts, no positive survives the pre-filter, so the reranker in Dense + Rerank can never score one. This illustrates our hypothesis on why Dense + Rerank performs worse in Section 4.1. The exact effect of this is still open to further ablation experiments

Table A3 gives the detailed results of retrieval systems performed on each of the MPs.

A one-way analysis of variance (ANOVA) was conducted to examine differences in LLM-as-judge scores across the three buckets. The results revealed a statistically significant effect of group, $F(2, 137) = 29.44$, $p < .001$, $\eta^2 = .30$, indicating a large effect. Post-hoc comparisons are presented in Table A4. Bucket C ($M = 1.35$, $SD = 0.78$) scored significantly lower than both Bucket A ($M = 3.79$, $SD = 1.33$) and Bucket B ($M = 3.51$, $SD = 1.43$), both $p < .001$. The difference between Bucket A and Bucket B was not statistically significant ($p = .295$), supporting the results from Section 3.4. We note that the 1–5 judgements are ordinal and nested within labels, whereas this analysis treats them as independent continuous observations; the bucket separation is large enough that this simplification is unlikely to change the qualitative conclusion, but a mixed-effects ordinal model would be more appropriate.

ID	Name of the World Heritage MP	Short Name	#Chunks	#Labels	#Lab. Chunks	#Pairs	Year
$\mathcal{H}_0$	*Colonies of Benevolence*	*Colonies*	847	11	30	43	2020
$\mathcal{H}_1$	*Frontiers of the Roman Empire – The Lower German Limes*	*LGL*	313	12	32	45	2021
$\mathcal{H}_2$	*Eise Eisinga Planetarium in Franeker*	*Eisinga*	334	10	32	43	2021
$\mathcal{H}_3$	*Dutch Water Defence Lines*	*Defence Lines*	525	13	14	32	2018
$\mathcal{H}_4$	*The Wadden Sea*	*Wadden Sea*	203	21	125	233	2023
$\mathcal{H}_5$	*The seventeenth-century canal ring area of Amsterdam inside the Singel-gracht*	*Amsterdam*	338	13	27	41	2023
$\mathcal{H}_6$	*Van Nellefabriek*	*Van Nelle*	48	4	5	6	2021
$\mathcal{H}_7$	*Schokland and surroundings*	*Schokland*	28	8	17	34	2019
$\mathcal{H}_8$	*Mill Network at Kinderdijk-Elshout*	*Kinderdijk*	143	13	19	28	2022
$\mathcal{H}_9$	*Droogmakerij de Beemster*	*Beemster*	—	—	—	—	2022
$\mathcal{H}_{10}$	*Rietveld Schröder House*	*Rietveld*	147	10	16	24	2024
Total (excl. $\mathcal{H}_9$)			**2,926**	**24**	**317**	**529**	

Table A2: The 11 Dutch World Heritage MPs used in this study. #Chunks: Docling text chunks (with the default setting of chunk_size= 1,000, overlap= 200); #Labels: distinct strategy labels annotated in that MP (from $\mathcal{D}_1$) being evaluated; #Lab. Chunks: number of chunks that get assigned at least one label; #Pairs: unique positive label-chunk pairs via alignment ψ. Total #Pairs= 529 corresponds to the ground truth used in retrieval evaluation. $\mathcal{H}_9$ (Beemster) is excluded because it is the only Dutch-language MP. The links to all the PDFs of the MPs for the Dutch WH properties are provided as hyperlinks, accessed on 22 September, 2026. The years when the current collection of the MPs is published are also provided, as a reference to potential research on temporal development of policy corpora across timeframes.

MP	BM25	Bi-enc	D+Rnk	CE
Rietveld	0.25	0.21	0.23	**0.35**
Schokland	**0.39**	**0.37**	**0.37**	0.32
Wadden Sea	0.20	0.22	0.21	0.24
Kinderdijk	0.14	0.20	0.23	0.19
Eisinga	0.14	0.18	0.18	0.18
LGL	0.17	0.12	0.16	0.15
Colonies	0.16	0.06	0.08	0.13
Defence Lines	0.10	0.08	0.08	0.12
Amsterdam	0.16	0.07	0.07	0.12
Van Nelle	0.10	0.10	0.10	0.09

Table A3: Per-MP MAP score by system (sorted by Cross-encoder MAP descending). Bi-enc means Bi-encoder; D+Rnk means Dense + Rerank; and CE means Cross-encoder.

Comparison	ΔM	t	p	g
A vs. B	0.27	1.05	.295	0.19
A vs. C	2.44	9.64	$< .001$	2.04
B vs. C	2.17	9.20	$< .001$	1.65

Table A4: Pairwise post-hoc comparisons (Welch's t-test) for the LLM-as-judge scores across the buckets. g shows the Hedges' g effect size.

A.3 LLM-as-Judge Prompt and Results

When making the buckets, 4 samples are missing from the intended 2 from Bucket A, 3 from Bucket B, and 1 from Bucket C. Two labels fall short of this intention: *Research gap* has only one ground-truth pair in total and therefore contributes a single Bucket A item, and *Water management* is missing two Bucket B items and one Bucket C item due to dropped TULIP API calls.

The LLM judge is run at temperature 0 with the following template, instantiated per label-chunk pair. Chunk text is truncated to 1,500 characters. Responses are parsed as JSON; the `reason` field is retained for qualitative inspection but not used in any quantitative result. Table A5 shows qualitative examples of the LLM-as-judge evaluation across three sampling buckets for two climate-adaptive management strategy labels together with the LLM reasoning.

```
You are an expert in heritage management and
climate adaptation.

Heritage concept label: {label}
Definition: {definition}

Text chunk:
"""{chunk_text}"""

Task: Rate how relevant the text chunk is to
the heritage concept label above, on a scale of
1-5:
1 = not relevant
2 = slightly relevant
3 = moderately relevant
4 = relevant
5 = highly relevant

Respond with ONLY JSON, no other text: {"score":
<1-5>, "reason": "<one short sentence>"}
```

Bucket	Label	Score	Chunk excerpt	LLM reasoning
A (labelled with ground truth)	*Adaptation*	5	*"Integrate and mainstream the Trilateral Climate Change Adaptation Strategy into all activity fields... Continue monitoring the Trilateral Climate Change Adaptation Strategy and embed the results in long-term trilateral climate change policies."*	Explicitly implements and monitors a Climate Change Adaptation Strategy.
B (unlabelled top-ranked)	*Adaptation*	5	*"Climate change adaptation requires the ability to anticipate sharp fluctuations in water supply and deal with heat stress. To this end, dykes are strengthened and water retention sites built, including in the World Heritage Site."*	Describes specific adaptation measures; not in ground truth, yet clearly relevant.
B (unlabelled top-ranked)	*Assessment*	5	*"assessed in terms of recognisability and authenticity: 1. Attribute possesses authenticity and integrity, and gives a good and perceptible impression of how the attribute looked in the historical context in the period concerned. 2. Attribute illustrates what it was historically, but the historical context has not been well preserved. 3. At[tribute...]"*	The text explicitly outlines a structured assessment framework to evaluate the authenticity and recognisability of heritage attributes.
B (unlabelled top-ranked)	*Assessment*	4	*"Spacial quality advisory team (as a complement to the ICOMOS guidelines), have drawn up recommendations regarding situations in which a Heritage Impact Assessment is desirable, and on formulating the terms of reference of future Heritage Impact Assessments and how to implement them. Over recent years, an extensive network of participating part[ners...]"*	The text explicitly mentions the use and implementation of Heritage Impact Assessments to evaluate impacts on the site.
C (unlabelled low-ranked)	*Adaptation*	1	*"The roads are damaged by speeding cars as well as by heavy (agricultural) traffic. Walkers, cyclists and horse riders sometimes come into conflict with the motorised traffic."*	Traffic management only; no climate content.

Table A5: Qualitative examples of LLM-as-judge evaluation for two labels. Bucket A shows a confirmed ground-truth positive (score 5). Bucket B shows a few high-ranked chunks not in the ground truth that the LLM judges equally relevant (score 5 or 4) as potential candidates of false negatives. Bucket C shows a low-ranked chunk the LLM correctly judges irrelevant (score 1). The column of Chunk excerpt is subsets of the complete 1000-character chunks selected by the evaluating LLM, and the column of LLM reasoning records its judgement criteria.

ESG-CID-Plus: A Stratified Benchmark for Disclosure Content Index Table Extraction from Corporate Sustainability Reports

Shafiuddin Rehan Ahmed, Reza Abdi, Tianyu Luan
Teja Kanchinadam, Christopher Clarke, Eugene Siow
Agentic AI Center of Excellence, Accenture, Mountain View, CA, USA
{shafiuddin.r.ahmed,c.a.clarke,eugene.siow}@accenture.com

Abstract

Corporate sustainability reports include a *content index* mapping each disclosure required by a reporting framework (e.g. GRI) to its supporting pages — a resource that potentially benefits nearly every downstream sustainability-NLP task, yet no existing benchmark scores its recovery as a framework-grounded, page-resolved output. We formalize Content Index Detection (CID) as three evaluable stages — index page retrieval (S1), table extraction and disclosure linking (S2), and page-label-to-index resolution (S3) — and release `esg_cid_plus`: 37 GRI reports and 7,294 rows stratified across three difficulty tiers, from unchanged real reports to an adversarial tier of naturally hard documents and synthetic perturbations. Our main benchmark sweeps open-weight vision-language models from sub-1B to 35B parameters (plus frontier-API references) as S2 extractors under shared, deterministic S1 and S3 stages: the best backends reach S2-F1 of 0.89–0.92 on the easier tiers, while standard OCR-benchmark scores explain only a third of the variance in extraction quality. Composing extraction with the deterministic resolver exposes S3 as the open frontier: under label-scheme resolution alone, where even best extraction caps end-to-end F1 at 0.11 on the hardest tier. `esg_cid_plus` will be released at https://huggingface.co/datasets/ahmeshaf/esg_cid_plus.

1 Introduction

Corporate sustainability reports (CSRs) are the primary public record of how organisations disclose their Environmental, Social, and Governance (ESG) practices. They are long, narrative-heavy PDFs whose structure is dictated by reporting frameworks such as the Global Reporting Initiative (GRI), the European Sustainability Reporting Standards (ESRS), the Sustainability Accounting Standards Board (SASB), or the recommendations of the Task Force on Climate-Related Financial Disclosures (TCFD).[1] Most CSRs also include, often as an appendix, a *content index*: a table that maps each required disclosure (e.g. GRI `302-3`, "Energy intensity") to its corresponding pages.

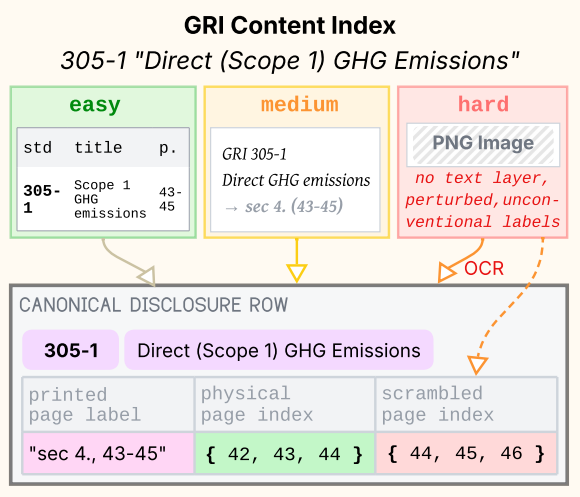


Figure 1: One canonical disclosure (GRI `305-1`, "Direct (Scope 1) GHG Emissions") rendered three different ways across the three difficulty tiers (`gri_easy`, `gri_medium`, `gri_hard`) must collapse to the same canonical row $r_i = (s_i, t_i, \tilde{t}_i, \ell_i, P_i)$. The footer separates the printed page label ℓ_i from the resolved 0-based PDF indices P_i; the dashed arrow on `gri_hard` marks where text-only extraction fails and OCR is required.

This index, the report's own disclosure-to-evidence map, is a resource that potentially benefits nearly every downstream sustainability-NLP application: retrieval-augmented generation (Ahmed et al., 2025; Ni et al., 2023), coverage assessment (Usmanova et al., 2025; Polignano et al., 2022), claim verification and grounded QA over CSRs (Vaghefi et al., 2023; Diggelmann et al., 2020), and longitudinal comparison of disclosure quality (Schimanski et al., 2024b; Ong et al., 2025). Yet the literature invests heavily in downstream disclo-

[1] See https://www.globalreporting.org/ (GRI); https://www.efrag.org/ (ESRS); https://sasb.org/ (SASB); and https://www.fsb-tcfd.org/ (TCFD).

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 183–198
October 28, 2026 ©2026 Association for Computational Linguistics

sure analysis while assuming the content index can already be recovered reliably.

We study *Content Index Detection (CID)*: given a CSR PDF $\mathcal{D}$ and a target framework $\mathcal{F}$, identify the pages $P_{\mathcal{F}}$ containing $\mathcal{F}$'s content index and produce a schema-aligned table $\mathrm{CI}_{\mathcal{F}}(\mathcal{D})$ with one row per framework disclosure $(s_i,\ t_i,\ \tilde{t}_i,\ \ell_i,\ P_i)$— the framework id and title, the title as it appears in the report, the verbatim page-reference cell text, and the resolved 0-based PDF page indices (Figure 1). We decompose CID into three stages: CI page retrieval (S1), CI table extraction (S2, comprising raw row extraction S2.1 and disclosure linking S2.2), and page reference resolution (S3) (Section 3), enabling systems to be evaluated by failure stage rather than only end-to-end.

Framed this way, CID is a *composition* of five fundamental document extraction tasks, each with its own literature and failure modes: (i) **Search** — retrieving the few index pages within a long report (S1); (ii) **Parse** — extracting table rows, including OCR and multi-page stitching (S2.1); (iii) **Bind** — disambiguating repeated *management-of-material-topics* entries (e.g., "3-3") by binding each to its spanning section header; (iv) **Fill** — mapping the parsed, disambiguated rows into the framework template, one canonical slot per disclosure (S2.2); and (v) **Resolve** — converting human-facing page references to physical 0-based PDF indices (S3). Prior work studies these tasks in isolation or, in ESG-NLP, uses CID only to bootstrap retrieval supervision (Ahmed et al., 2025); no benchmark evaluates them jointly on a page-resolved, framework-grounded output, and *Bind* and *Resolve* remain essentially unaddressed.

Two of these tasks make CID markedly harder than generic table extraction (Mathew et al., 2021; Saad-Falcon et al., 2023; Schimanski et al., 2026). The index table is idiosyncratic across CSRs — variable column layouts, multi-page two-column spreads, and disclosure ids in many typographic variants ("GRI 302-3", "302–3", "GRI 302:3") — and its page references are written for human readers as *printed page labels* ("iv", "A-12", "2.7") that need not match 0-based PDF indices: numbering schemes restart or switch styles, the same label can recur across physical pages, and a reference may even be a section name or an external citation ("Annual Report 2023, p. 12"), so resolution must model the report's own label scheme — or correctly mark a reference unresolvable. These failure modes are also not uniform: heuristics that work on one report family break on another, so a benchmark for CID must stratify reports by difficulty — from unchanged real reports, through naturally adversarial ones, to synthetically perturbed documents — to expose where a single recipe fails and why.

Concretely, we make four contributions.

Task formalisation. We formalise CID as a framework-grounded, page-resolved problem over evaluable stages S1–S3 (Section 3).

Stratified benchmark. We release the dataset `esg_cid_plus`: 37 GRI reports and 7,294 CID rows across three tiers — `gri_easy`, `gri_medium`, and an adversarial `gri_hard` (§4).

Baselines. We implement a deterministic baseline for every stage — lexical S1 retrieval with OCR fallback, Img2Table+RapidOCR S2 extraction and linking, and a label-scheme S3 resolver — as fixed, swappable components (§5).

An S2 model sweep. We sweep open-weight vision-language models from sub-1B to 35B parameters (plus frontier-API references) as S2.1 extractors under oracle S1, so that extraction quality is isolated (§6): the best backends reach S2-F1 of 0.89–0.92 on the easier tiers but at most 0.79 on `gri_hard`, and standard OCR-benchmark scores explain only a third of the variance in extraction quality (§7). An ablation separating standard from topic-qualified identifiers isolates the cost of *Bind* as standing headroom.

Our results indicate that page resolution (S3) is the open frontier: under deterministic resolution, even gold-quality extraction caps end-to-end F1 at 0.11 on `gri_hard` (§6) — the remaining headroom sits in S3. `esg_cid_plus` thus lays the foundation for future work that models *Resolve* directly rather than treating page resolution as a deterministic post-process.

2 Related Work

We position `esg_cid_plus` against three bodies of work — table structure recognition, long-document VQA and key-information extraction, and ESG-NLP — and map each onto the five tasks CID composes (§1), as summarised in Table 1.

Table structure recognition. FinTabNet (Zheng et al., 2021) and PubTables-1M (Smock et al., 2022b) provide large-scale training signal for *Parse*, the latter with explicit spanning-cell classes,

Dataset / Benchmark	Document Extraction Tasks (§1)				
	Pg. Retr.	Table Understanding			Pg. X-Ref.
	Search	Parse	Bind	Fill	Resolve
Table structure recognition					
FinTabNet	✗	✓	▲	✗	✗
PubTables-1M	✗	✓	▲	✗	✗
HiTab	✗	▲	✓	▲	✗
TUTA	✗	▲	✓	▲	✗
Long-document VQA and KIE					
DocVQA	▲	▲	✗	✗	✗
MP-DocVQA	✓	▲	✗	✗	✗
Kleister	▲	✗	✗	✓	✗
DocILE	✗	✓	✗	✓	✗
ESG / sustainability benchmarks					
ESGDoc	✓	▲	✗	✗	▲
GRI Pipeline	▲	▲	✗	▲	✗
ReportGRI	▲	✗	✗	✓	✗
GRI-QA	▲	✓	✗	✓	✗
`esg_cid`	▲	✓	✗	▲	✗
ours	✓	✓	✓	✓	✓

Table 1: Coverage of the five document extraction tasks CID composes (§1) across the three bodies of prior work surveyed below — table structure recognition, long-document VQA and key-information extraction, and ESG / sustainability benchmarks. ✓ explicitly evaluated with released gold; ▲ present in the data or method but not scored; ✗ absent. Citations are given in the text.

and GriTS (Smock et al., 2022a) scores recovered topology, content, and location; Nunes et al. (2025), Soric et al. (2025), and Horn and Keuper (2026) benchmark multimodal LLMs against OCR pipelines on PDF tables, and OCRBench (Liu et al., 2024) and OCRBench v2 (Fu et al., 2025) rank VLMs on OCR-centric recognition, localisation, and parsing — they guided our open-weight roster selection but score isolated capabilities. Closest to *Bind*, HiTab (Cheng et al., 2022) and TUTA (Wang et al., 2021) formalise linking a data cell to its governing multi-level headers, but both assume the table is already parsed, bind cells to *in-table* headers rather than an external framework schema, and never recover spans from a rendered PDF. GRI content indices additionally span multiple physical pages as one logical table, so per-fragment structural metrics are uninterpretable; we capture location-text quality within S2.2 via fuzzy matching instead.

Long-document VQA and key-information extraction. DocVQA (Mathew et al., 2021) and PDFTriage (Saad-Falcon et al., 2023) show that modelling document structure beats flat-text retrieval, motivating our staged decomposition, and MP-DocVQA (Tito et al., 2023) makes evidence-page identification over multi-page documents explicit (*Search*) — but the located page is evidence for an answer, not a resolved output key. On the extraction side, Kleister (Stanisławek et al., 2021) extracts values for a fixed key schema from long PDFs and DocILE (Šimsa et al., 2023) localises line items on invoices (*Fill*); both fill flat key schemas, without hierarchical-header binding or per-row page resolution.

ESG and sustainability NLP. The field has grown quickly — domain-adapted language models (Webersinke et al., 2022; Schimanski et al., 2024b), retrieval-augmented report QA (Ni et al., 2023; Vaghefi et al., 2023), and climate-claim verification (Diggelmann et al., 2020; Schimanski et al., 2024a; Mankour et al., 2025) — but none targets framework alignment or is benchmarked page-resolved, though CID's schema would readily sharpen their claims. On framework alignment specifically: Usmanova et al. (2025) automate GRI alignment via zero-shot prompting (five reports, no tables, no page grounding); Polignano et al. (2022) apply an NLP pipeline to GRI indexes across 134 reports but release no gold; GRI-QA (Contalbo et al., 2025) benchmarks table QA over disclosed values once located — the natural downstream complement to CID. ESGDoc (Wang et al., 2023) is the closest structural predecessor, extracting a report's own chapter hierarchy rather than the compliance-framework table, with no label-to-index pagination — leaving *Resolve* without a dedicated benchmark anywhere.

Positioning. `esg_cid_plus` is the only benchmark in Table 1 scoring all five tasks jointly, so an end-to-end failure can be attributed to a stage: pages missed (*Search*), rows lost (*Parse*), topics unbound (*Bind*), slots mis-filled (*Fill*), or references unresolved (*Resolve*). The closest prior work, `esg_cid` (Ahmed et al., 2025), mines content indices to bootstrap retrieval supervision and supplies the PDFs for the bulk of our adapted subset, but collapses duplicate topic-qualified rows into a flat namespace and releases no page-resolved gold, so neither *Bind* nor *Resolve* can be measured on it.

3 Content Index Extraction Task (CID)

Before turning to dataset construction (§4), we formalize CID.

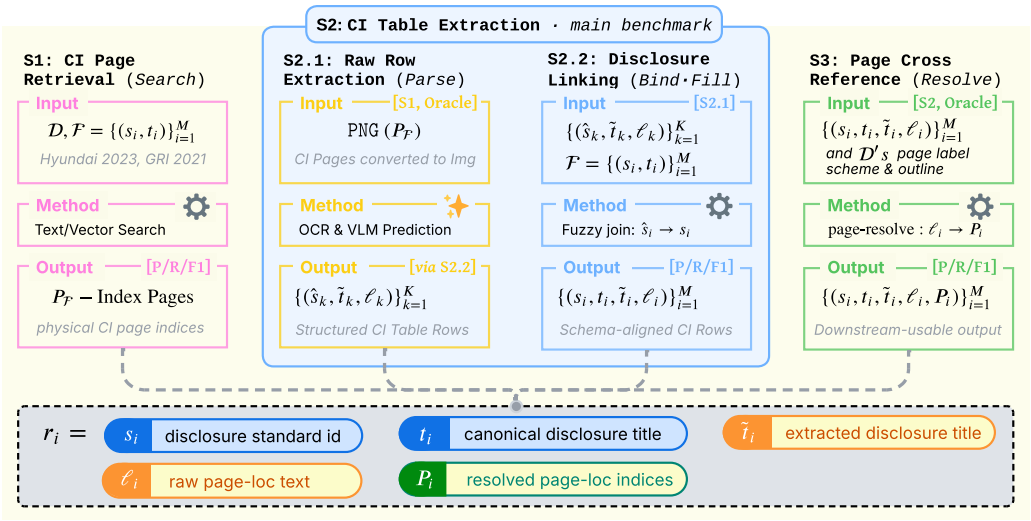


Figure 2: The three stages of CID extraction. Given a report $\mathcal{D}$ and a framework $\mathcal{F}$: (S1) retrieves index pages $P_{\mathcal{F}}$ (*Search*); (S2) extracts raw candidate rows (S2.1, *Parse*) and fuzzy-joins them to the canonical schema (S2.2, *Bind* · *Fill*); (S3) resolves verbatim page-location text ℓ_i to physical page indices P_i via $\mathcal{D}$'s printed-label scheme (*Resolve*). Icons mark each stage as model-inferred (✨) or deterministic (⚙). S2.1 is the only inferred stage and the target of our main benchmark.

3.1 Task Formalization

Given a PDF report $\mathcal{D}$ (N pages) and a disclosure framework $\mathcal{F} = \{(s_i, t_i)\}_{i=1}^{M}$ — a canonical list of standard ids (e.g. `"2-1"`) paired with titles (e.g. `"Organizational details"`), such as GRI 2021 (M=147) — CID requires producing two outputs. $P_{\mathcal{F}} \subseteq \{0, \ldots, N-1\}$ is the set of PDF pages on which $\mathcal{F}$'s content index appears. $\mathrm{CI}_{\mathcal{F}}(\mathcal{D})$ is the extracted, schema-aligned content index:

$$\mathrm{CI}_{\mathcal{F}}(\mathcal{D}) = \{r_i : i = 1, \ldots, M\},$$

with exactly one row per disclosure:

$$r_i = (s_i,\ t_i,\ \tilde{t}_i,\ \ell_i,\ P_i).$$

$\tilde{t}_i$ is the title as printed in the report (may differ from canonical t_i due to abbreviation, translation, or OCR noise); ℓ_i is the verbatim location-reference cell (numeric labels, e.g. `"62, 78"`; section names, e.g. `"Governance chapter"`; external citations, e.g. `"Annual Report 2023, p. 12"`; or inline notes, e.g. `"Not applicable"`); $P_i \subseteq \{0, \ldots, N-1\}$ is the resolved set of physical page indices. When a report provides no evidence for s_i within $\mathcal{D}$, P_i=$\emptyset$. $\mathrm{CI}_{\mathcal{F}}(\mathcal{D})$ is *M-complete*: every $s_i \in \mathcal{F}$ has a row in $\mathrm{CI}_{\mathcal{F}}(\mathcal{D})$.

3.2 Subtasks

The task decomposes into three stages illustrated in Figure 2, with stage S2 further divided into two sub-stages.

S1: CI Page Retrieval (*Search*). Given $(\mathcal{D}, \mathcal{F})$, produce $P_{\mathcal{F}}$ — the set of pages on which the content index appears. This is a focused retrieval problem: the index shares surface vocabulary with tables of contents, assurance statements, and appendix tables.

S2: CI Table Extraction. Given $(\mathcal{D}, \mathcal{F}, P_{\mathcal{F}})$, produce the M-complete linked table in two sub-stages:

S2.1: Raw Row Extraction (*Parse*). Extract raw candidate rows $\{(\hat{s}_k, \tilde{t}_k, \ell_k)\}_{k=1}^{K}$ from the index region, handling OCR, multi-page stitching, and multi-column layouts.

S2.2: Disclosure Linking (*Bind* · *Fill*). Fuzzy-join candidates against $\mathcal{F}$ on $\hat{s}_k \rightarrow s_i$: candidates with no canonical match are dropped, duplicates are merged, and schema entries with no candidate become ε rows, yielding the M-complete table $\{r'_i\} = \{(s_i, t_i, \tilde{t}_i, \ell_i)\}$.

S3: Page Cross Reference (*Resolve*). Given each r'_i and $\mathcal{D}$, resolve $\ell_i \rightarrow P_i$ by recovering the report's label-to-index function λ and resolving each token by type: numeric labels via λ^{-1}, section references via GoTo links or heading lookup, external citations to P_i=$\emptyset$.

We now turn to how this formalism is annotated, stratified, and evaluated in practice.

4 Dataset Construction

`esg_cid_plus` is built from the 73 GRI PDFs in `airefinery/esg_cid` (57 `train`, 6 `dev`, 10 `test_gri`), covering GRI Standards 2016 (M=238) and GRI Standards 2021 (M=147); the 11 ESRS documents are set aside as a separate future task. We select 26 of the 73 reports for three real-world splits, stratified by page-label scheme complexity and index layout, and annotate ten additional reports for an adversarial hard tier. The full per-report listing is in Appendix A.

Annotation protocol. The annotator starts from the framework template $\mathcal{F}$ and, per PDF: (i) locates the content index pages, recorded as the S1 ground truth; (ii) for each $s_i \in \mathcal{F}$, copies the location cell text ℓ_i verbatim from the matching content-index row (left empty if none exists); (iii) resolves each populated ℓ_i to physical page indices P_i, following section-name pointers where needed.

Row taxonomy. Every row of $\mathrm{CI}_{\mathcal{F}}(\mathcal{D})$ is first *absent* ($\ell_i{=}\varepsilon$, count M^{abs}, inserted by step (ii) to enforce M-completeness) or *mentioned* ($\ell_i{\neq}\varepsilon$, count M^{men}; gold target for S2), so $M = M^{\mathsf{abs}} + M^{\mathsf{men}}$. Mentioned rows further split into *page-indexed* ($P_i{\neq}\emptyset$, count M^{pg}; gold target for S3 — numeric spans and resolvable section pointers) and *non-page-indexed* ($P_i{=}\emptyset$, count $M^{\overline{\mathsf{pg}}}$; omission markers, external references, unresolvable section pointers), so $M^{\mathsf{men}} = M^{\mathsf{pg}} + M^{\overline{\mathsf{pg}}}$. Per-split totals are unpacked in Section A.3.

Standard vs. topic-qualified disclosures. Most disclosures are *standard* (e.g. `2-1`): one row, one identity. Material-topic disclosures (`103-1/2/3` under GRI 2016, `3-3` under GRI 2021) instead recur once per material topic under an identical ID and title — the same binding problem as a data cell under a multi-level header (§2), except the governing header sits in the surrounding section, not the table. GRI 2016 crosses 96 such disclosures against up to 32 topics; GRI 2021 has 32. We evaluate baselines on the standard disclosures and analyze topic-qualified rows separately.

Difficulty tiers. Tiers are defined by construction-side signals observable before any model is applied. We classify each PDF's label-to-index function λ as *trivial* (declared and machine-readable), *nontrivial* (must be inferred empirically, e.g. from sampled headers), or *complex* (requires parsing document structure). **`gri_easy`** (10 PDFs): trivial scheme, contiguous well-labelled index. **`gri_medium`** (10 PDFs): a nontrivial scheme (6/10), sidebar/multi-column layout, or non-standard headers. **`gri_hard`** (10 PDFs): six real-world adversarial reports (section-based location text, double-page-spread labeling, or a complex hierarchical scheme) plus four synthetic `_x` perturbations (image-only rendering or scrambled `/PageLabels`) applied to easy/medium source reports; Table 4 (Appendix A) lists all ten.

5 Experimental Setup

Of the four stages formalized in §3.2 (Figure 2), only S2.1 is model-inferred (✨); S1, S2.2, and S3 are deterministic (⚙) and shared by all backends, so end-to-end differences reflect the S2.1 backend alone.

5.1 Baselines

S1: CI Page Retrieval (*Search*). We index each page with `lunr` (Diaz, 2018), a lightweight TF-IDF lexical search engine, and query it using batched disclosure ids and titles from $\mathcal{F}$. We take the top-10 retrieved pages as seeds and expand each by a $\pm$5-page neighborhood. We then confirm candidate pages using *disclosure-id density*: we normalize framework-id regex matches by token count and accept a page if its density exceeds 0.03 or, for sparse pages, if it contains at least seven matches regardless of density. We tuned the five retrieval hyperparameters (seed count, neighborhood radius, density threshold, match-count floor, and the OCR fallback below) once on `train_small` (S1 P/R/F1 = .95/.98/.96) and kept them fixed across all evaluation splits. Finally, we close confirmed pages under contiguity by filling gaps between adjacent detections. When a page has an empty or sparse text layer, we transparently substitute **RapidOCR** (RapidAI, 2021) output before indexing and density scoring, allowing the same retrieval pipeline to handle rasterized reports without a separate code path.

S2: CI Table Extraction (*Parse · Bind · Fill*). We render confirmed CI pages as images and parse them with **Img2Table** (Canton, 2021)+**RapidOCR** (S2.1). We scan each extracted row for a disclosure id: an id match opens a new candidate row, while subsequent rows without an id extend the current candidate by appending their title and location-text cells until the next id match. This recovers disclosures whose title or location text spans multiple

Model	S2-F1	Model	S2-F1	Model	S2-F1
GPT-5.5	0.85	Qwen3.5-35B-A3B	0.65	Ministral-3-3B	0.48
GPT-5.4	0.80	Nemotron-Nano-v2-VL	0.65	Gemma-4-12B-it	0.45
GPT-5-mini	0.80	Ministral-3-8B	0.64	Nemotron-3-Nano-Omni	0.45
Qwen3.5-27B	0.79	Ovis2.5-9B	0.63	LLaVA-OneVision-2	0.42
Gemini-3.5-Flash	0.79	Claude Haiku 4.5	0.61	UI-TARS-1.5-7B	0.34
Gemini-3.1-Flash-Lite	0.78	GPT-5-nano	0.60	InternVL3.5-2B	0.31
Gemini-3-Flash	0.77	Qwen3.5-9B	0.59	Gemma-4-E4B-it	0.31
Qwen3.6-27B	0.77	MiniCPM-V-4.5	0.58	Qwen3.5-2B	0.28
Ovis2.6-30B-A3B	0.74	InternVL3.5-4B	0.56	Granite-Vision-4.1-4B	0.28
Qwen3.6-35B-A3B	0.74	InternVL3.5-8B	0.56	InternVL3.5-1B	0.27
Gemma-4-31B-it	0.71	Qwen3.5-4B	0.56	MiniCPM-V-4.6	0.23
Claude Sonnet 4.5	0.71	**Baseline (Img2Table)**	**0.52**	Ovis2.5-2B	0.21
GLM-4-6v-flash	0.69	Qianfan-OCR	0.52	Gemma-4-E2B-it	0.19
Gemma-4-26B-A4B-it	0.67	Step3-VL-10B	0.51	Qwen3.5-0-8B	0.16
olmOCR-2-7B-1025	0.66	InternVL3.5-14B	0.50	Ovis2-1B	0.13
Ministral-3-14B	0.66	InternVL3.5-30B-A3B	0.50	Granite-Vision-3.2-2B	0.06

Table 2: Mean S2-F1 across `gri_easy`, `gri_med`, and `gri_hard` (oracle S1). Closed-source models are indicated in blue; the Baseline (§5.1) is indicated in red.

table rows. We then normalize candidate ids (removing whitespace, dash variants, and framework prefixes) and join them to $\mathcal{F}$ by exact id match (S2.2). We discard candidates with no schema match and emit ε rows for schema disclosures with no matching candidate. Every S2.1 backend shares this normalize-then-join stage, so it absorbs backend-specific formatting noise identically before scoring.

S3: Page Cross Reference (*Resolve*). We read the PDF's `/PageLabels` catalog (numbering style, prefix, and start page) and build a label→physical-index map, defaulting to implicit arabic-from-zero numbering when the catalog is absent. We then mine each row's location text for numeric ranges, bare numbers, and roman or alphabetic labels after scrubbing disclosure ids (so, e.g., `"2-1"` is not misread as a page range). Finally, we resolve each mined label through the map to produce P_i, dropping labels that do not appear in the map. This explains why non-page-indexed mentioned rows (§4) exist even when ℓ_i is populated.

5.2 Metrics

We report three canonical metrics, one per stage boundary. **S1-P/R/F1** is macro P/R/F1 over per-report page tuples $(\mathcal{D}, p)$, $p \in P_{\mathcal{F}}$. **S2-P/R/F1** is scored at the end of S2.2: micro P/R/F1 over *mentioned* tuples $(\mathcal{D}, s_i, \ell_i)$ (§4), where a true positive requires both the correct disclosure identifier s_i and the extracted verbatim location text $\hat{\ell}_i$ to fuzzy-match the gold ℓ_i at $\theta \geq 85$ (`rapidfuzz` token-sort ratio). **S3-P/R/F1** is the end-to-end score after S3: micro P/R/F1 over $(\mathcal{D}, s_i, p)$ tuples, one tuple per resolved page index $p \in P_i$; a predicted tuple is a true positive only if both s_i and that specific physical page index p exactly match a gold tuple.

Topic-qualified identifiers (*Bind*). Gold topic-qualified rows (e.g. `3-3:Materials`, §4) are *included* in all reported scores. No evaluated system attempts topic binding — the deterministic extractor merges recurring `3-3` rows into one candidate, and the VLM prompt requests unqualified ids — so these gold rows are unreachable for every backend and uniformly bound recall. Scoring each configuration with and without topic-qualified identifiers therefore isolates the cost of *Bind* as a controlled ablation.

5.3 S2 Model Sweep Setup

For the sweep, the same rendered CI page images are given directly to a vision-language model prompted for the same (id, title, location-text) triples — no table-parsing step. S2.1 is the only stage swept across backends; every backend is composed with oracle S1 and the shared S2.2/S3 stages above (Table 2). The sweep evaluates 48 backends under one shared harness: open-weight VLMs are self-hosted on Kubernetes (AKS) with H100 GPU nodes, served via vLLM (Kwon et al., 2023); closed models are queried through their provider APIs behind the same gateway. Appendix B gives further detail.

Roster curation. We used OCRBench (Liu et al., 2024) — the OCR-centric benchmark most consistently reported across candidate models' technical reports and model cards — as a guide: back-

Split	Gold	S1			S2			S3		
		P	R	F1	P	R	F1	P	R	F1
gri_easy	all	.93	.97	.93	.64	.51	.56	.73	.56	.64
	w/o topic	(same)			.64	.53	.58	.73	.67	.70
gri_med	all	.93	.96	.92	.73	.57	.64	.69	.46	.55
	w/o topic	(same)			.73	.59	.65	.69	.58	.63
gri_hard	all	.92	.78	.79	.36	.29	.32	.26	.06	.10
	w/o topic	(same)			.36	.30	.33	.26	.08	.12

Table 3: Deterministic baseline scored at each stage boundary, with S2/S3 composed on the baseline's own detected S1. **all** scores against the full gold; **w/o topic** excludes topic-qualified identifiers — the *Bind* ablation (S1 is unaffected).

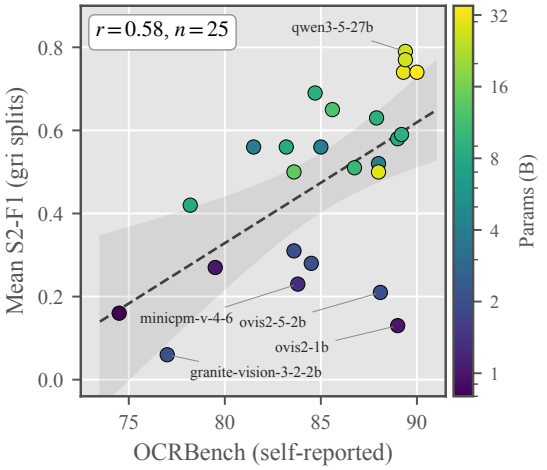


Figure 3: Mean S2-F1 vs. self-reported OCRBench (Liu et al., 2024), for the 25 open-weight backends with a comparable score. Points are colored by parameter count (total for mixture-of-experts); the dashed line is the OLS fit with an approximate 95% band.

ends were selected to span its full published score range, not just current leaderboard leaders, so the sweep covers varying document-reading competence rather than only strong OCR readers. Twenty-five of the evaluated open-weight models report a comparable OCRBench number; the frontier APIs and several families (e.g. the `ministral-3` models and most `gemma-4` variants) do not publish one, as do four models whose only available figure comes from a third-party competitor table — all excluded from the OCRBench-conditioned analysis of §7, which checks whether that guide actually predicted extraction quality. Per-backend references are listed in Appendix B.3.

6 Results

6.1 Deterministic Baseline

Table 3 reports the fully deterministic pipeline (§5) at each stage boundary on the three evaluation splits. S1 detection is near solved on `gri_easy`; the OCR fallback is what keeps `gri_hard` viable (without it, hard-split S1-F1 drops from .79 to .60). The S2 and S3 scores compose the baseline's own detected S1, so they inherit these S1 errors; `gri_hard` collapses at S3 (.10) — the page-resolution pathologies analysed below. The *Bind* ablation (§5.2) excludes topic-qualified identifiers from gold: S2-F1 moves by at most .02, but S3-F1 rises by .06/.08/.02 on the three splits — recall the baseline forfeits by never attempting topic binding.

6.2 Selecting the S2.1 Backend: A VLM Sweep

We sweep open-weight VLMs spanning sub-1B to 35B parameters (plus frontier API references) as S2.1 backends, all evaluated under oracle S1 with shared S2.2/S3 code. Unlike prior VLM-vs-OCR table-extraction benchmarks (Nunes et al., 2025; Soric et al., 2025; Horn and Keuper, 2026) (§2), which compare models on isolated single-page tables, GRI content indices routinely span multiple physical pages as a single logical table, so extraction quality here is measured by its effect on a composed, framework-grounded output.

Table 2 ranks all evaluated models by mean S2-F1 across the three splits, grouped by model family and ordered by parameter count (mixture-of-experts sizes shown as total/active).

7 Analysis

7.1 Relationship to Standard OCR Benchmarks

Figure 3 tests whether OCRBench, used to guide roster curation (§5.3), actually predicts S2 extraction quality, plotting mean S2-F1 against it for the 25 open-weight backends with a comparable score. The association is positive but loose: Pearson $r = 0.58$ ($p = 0.002$; Spearman $\rho = 0.61$), so OCRBench accounts for only about a third of the variance in S2-F1 ($r^2 = 0.34$). Reading competence is thus necessary but far from sufficient for content-index extraction.

The residuals are structured by model capacity: several sub-2B models (e.g. `ovis2-1b`, `qwen3-5-2b`) post a competitive OCRBench (77–89) yet collapse to S2-F1 below .25, and the signed residual increases with parameter count (Spearman $\rho = 0.60$). We read this as a task mis-

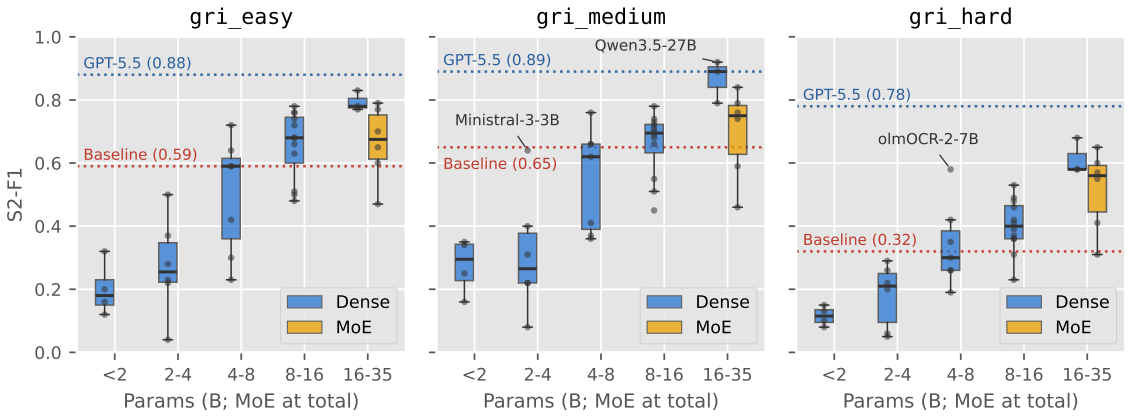


Figure 4: S2.1 location-extraction quality (S2-F1) by parameter bin, per split, for dense vs. mixture-of-experts open-weight VLMs under oracle S1 (closed models excluded, no published size); dotted lines mark GPT-5.5 and the deterministic Baseline. Boxes show median and IQR; dots are individual models. Annotated outliers: Ministral-3-3B, olmOCR-2-7B, and Qwen3.5-27B, the only open model to surpass GPT-5.5.

match, not noise: OCRBench rewards recognizing characters in single-region crops, already saturated at this size, while S2 demands multi-page, schema-faithful structured prediction and instruction following—these models can compose a table (many emit HTML natively), but our task's specific structuring, a multi-page (id, title, location-text) triple matched against $\mathcal{F}$'s fixed schema, breaks that flow. Other OCR-centric benchmarks we collected show the same trend (DocVQA (Mathew et al., 2021), $r = 0.68$; CharXiv (Wang et al., 2024), $r = 0.70$), so the pattern is not specific to OCRBench. Since these small models' character-recognition competence is already saturated, a natural line of future work is to repurpose them as pure OCR engines feeding the deterministic S2.1 pipeline (§5.1) — in place of RapidOCR — rather than asking them to also structure the output, and compare against the Img2Table+RapidOCR baseline directly.

7.2 S2 Extraction Quality

Figure 4 summarizes the S2.1 sweep (§6.2) as S2-F1 distributions over parameter bins, separating dense from mixture-of-experts models. Three patterns are consistent across splits. First, closed frontier models lead on the hard split (GPT-5.5 at .78 S2-F1 vs. .68 for the best open model), but the best open models are competitive or better on `gri_med` (Qwen3.5-27B at .92). Second, location-text fidelity scales with model capacity far more steeply than disclosure identification does: sub-4B models frequently extract the right rows but paraphrase or truncate location text, collapsing S2-F1 below .40. Third, split difficulty dominates model choice on `gri_hard`: every backend degrades substantially, consistent with the pagination and rasterization pathologies described above.

8 Conclusion

We formalized Content Index Detection (CID) as a composition of five document-extraction tasks — Search, Parse, Bind, Fill, and Resolve — spanning three evaluable stages, and released `esg_cid_plus`: 37 GRI reports and 7,294 M-complete rows stratified into `gri_easy`, `gri_med`, and an adversarial `gri_hard` tier, each row carrying both verbatim page-location text and resolved physical page indices. Our deterministic baseline and 35B-parameter VLM sweep converge on the same finding from opposite directions: page-location resolution (S3) is the binding bottleneck, not table extraction (S2.1). Even oracle-quality S2.1 rows only lift `gri_hard` S3-F1 to .11, and improvements in S2-F1 among strong backends stop translating into S3-F1 gains once extraction quality clears a moderate threshold — what predicts end-to-end success past that point is whether a backend's page references resolve cleanly, not how closely its location text matches the annotator's phrasing. This decomposition lets future work target *Resolve* directly — e.g. modelling a report's label scheme jointly with extraction, rather than treating page resolution as a deterministic post-process — and the stage-scored, page-resolved gold in `esg_cid_plus` makes that progress measurable rather than inferred from end-to-end scores alone.

Limitations

esg_cid_plus covers 37 GRI reports (10/10/10 across gri_easy/gri_med/gri_hard, plus 7 dev), substantially smaller than table-structure-recognition corpora such as PubTables-1M. Annotating page-resolved content indices requires a human to locate the index pages, transcribe location text verbatim, and resolve it against each report's own page-label scheme — a slow, per-report process that does not scale the way automatically mined tables do. This is most acute for gri_hard: ten reports is enough to surface the label-scheme and rasterization failure modes we report, but too few to estimate how S3 performance varies across the long tail of adversarial page-labeling conventions in the wild.

Ethics Statement

The corporate sustainability reports underlying esg_cid_plus are publicly disclosed by the companies themselves, contain no personal or sensitive information, and are used here strictly for their stated purpose: making ESG disclosures auditable. Annotation was carried out by the authors; no external crowd labor was used. We see this work as supporting, not substituting for, sustainability accountability: automating content-index recovery lowers the cost of checking whether a report actually discloses what a framework requires, which benefits auditors, investors, NGOs, and regulators assessing compliance at scale, and the stage-level errors we surface (particularly the S3 page-resolution bottleneck) are meant to caution against deploying automated disclosure-checking tools without accounting for where they still fail. We do not foresee a dual-use risk specific to this task: the same extraction pipeline that helps a third party verify a report's disclosures could in principle help a reporting company game a content index, but the dataset and models here only recover disclosure structure that a report already states, and do not evaluate the truthfulness of underlying ESG claims. esg_cid_plus will be released under the same license as its source PDFs' host repository, airefinery/esg_cid, at https://huggingface.co/datasets/ahmeshaf/esg_cid_plus.

Disclaimer

This content is provided for general information purposes and is not intended to be used in place of consultation with our professional advisors. This document may refer to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied.

Copyright © 2026 Accenture. CC BY-NC-ND. All rights reserved. Accenture and its logo are registered trademarks of Accenture.

References

Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. 2025. Enhancing retrieval for ESGLLM via ESG-CID: A disclosure content index finetuning dataset for mapping GRI and ESRS. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 1–16, Vienna, Austria. Association for Computational Linguistics.

AIDC-AI Ovis Team. 2026. Ovis2.6-30B-A3B: Mixture-of-experts multimodal large language model. Hugging Face model card, https://huggingface.co/AIDC-AI/Ovis2.6-30B-A3B. Official release artifact for the Ovis2.6 MoE VLM (30B total / 3B active parameters); no dedicated Ovis2.6 technical report exists.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, and 1 others. 2026. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*.

Anthropic. 2025a. Claude Haiku 4.5 system card. Accessed: July 2026.

Anthropic. 2025b. System card: Claude Sonnet 4.5. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf. Accessed: July 2026.

Xavier Canton. 2021. img2table: Table identification and extraction from PDFs and images. https://github.com/xavctn/img2table.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110. Association for Computational Linguistics.

Michele Luca Contalbo, Sara Pederzoli, Francesco Del Buono, Venturelli Valeria, Francesco Guerra, and Matteo Paganelli. 2025. GRI-QA: a comprehensive benchmark for table question answering over environmental data. In *Findings of the Association for*

Computational Linguistics: ACL 2025, pages 15764–15779, Vienna, Austria.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Guo Chen, and 1 others. 2025. NVIDIA Nemotron Nano V2 VL. *arXiv preprint arXiv:2511.03929*.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, and 1 others. 2026. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*.

Yeray Diaz Diaz. 2018. Lunr.py. https://github.com/yeraydiazdiaz/lunr.py. Python port of Lunr.js.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, and 1 others. 2026. Qianfan-OCR: A unified end-to-end model for document intelligence. *arXiv preprint arXiv:2603.13398*.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Google. 2025. Gemini 3 Flash: Frontier intelligence built for speed. https://blog.google/products/gemini/gemini-3-flash/. Model card: https://deepmind.google/models/model-cards/gemini-3-flash/. Accessed July 2026.

Google DeepMind. 2026a. Gemini 3.1 Flash-Lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/. Published March 3, 2026. Accessed July 2026.

Google DeepMind. 2026b. Gemini 3.5 Flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/. Published May 19, 2026. Accessed July 2026.

Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, and 1 others. 2025. Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence. *arXiv preprint arXiv:2502.09927*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Pius Horn and Janis Keuper. 2026. Beyond string matching: Semantic evaluation of PDF table extraction. *arXiv preprint arXiv:2603.18652*.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, and 1 others. 2026. Step3-VL-10B technical report. *arXiv preprint arXiv:2601.09668*.

IBM Granite Vision Team. 2026. Granite 4.1 Vision. Hugging Face model card, released April 29, 2026; accessed July 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, and 1 others. 2025. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*.

R. Mankour, Y. Chafai, H. Saleh, G. Ben Hassine, T. Barreau, and P. Tankov. 2025. Climate finance bench. https://github.com/Pladifes/climate_finance_bench. Working paper, PLADIFES.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

Guilherme Nunes, Vitor Rolla, Duarte Pereira, Vasco Alves, Andre Carreiro, and Márcia Baptista. 2025. Benchmarking table extraction: Multimodal LLMs vs traditional OCR. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 8–15, Vienna, Austria. Association for Computational Linguistics.

Keane Ong, Rui Mao, Ranjan Satapathy, Ricardo Shirota Filho, Erik Cambria, Johan Sulaeman, and Gianmarco Mengaldo. 2025. Explainable natural language processing for corporate sustainability analysis. *Information Fusion*, 115:102726.

OpenAI. 2026a. GPT-5.4 thinking system card. Published March 5, 2026; accessed July 2026.

OpenAI. 2026b. GPT-5.5 system card. Published April 23, 2026; accessed July 2026.

OpenBMB. 2026. MiniCPM-V 4.6. `https://huggingface.co/openbmb/MiniCPM-V-4.6`. Hugging Face model card. Accessed 2026-07-10.

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.

Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. olmOCR 2: Unit test rewards for document OCR. *arXiv preprint arXiv:2510.19817*.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. UI-TARS: Pioneering automated GUI interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents. Official release blog post; accessed July 2026.

Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model. Official release blog post. No Qwen3.6 technical report exists as of July 2026 (searched arXiv and Google Scholar; the QwenLM/Qwen3.6 GitHub README lists "Paper" as unlinked placeholder text and documentation as "coming soon"). Accessed July 2026.

Qwen Team. 2026c. Qwen3.6-35B-A3B: Agentic coding power, now open to all. `https://qwen.ai/blog?id=qwen3.6-35b-a3b`. Official release blog post. No Qwen3.6 technical report exists as of July 2026 (searched arXiv and Google Scholar). Cited as the closest official artifact for the evaluated checkpoint listed as "Qwen3.6-35B-A3B v2": no official v2/-2 variant of Qwen3.6-35B-A3B is documented by Qwen anywhere (Hugging Face, ModelScope, GitHub, or blog); only the single April 2026 release exists. Accessed July 2026.

RapidAI. 2021. RapidOCR: A lightweight, cross-platform OCR toolkit based on ONNX runtime. `https://github.com/RapidAI/RapidOCR`.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, David Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. 2023. PDFTriage: Question answering over long, structured documents. *Preprint*, arXiv:2309.08872.

Tobias Schimanski, Imene Kolli, Yu Fan, Ario Saeid Vaghefi, Jingwei Ni, Elliott Ash, and Markus Leippold. 2026. pdfQA: Diverse, challenging, and realistic question answering over PDFs. *Preprint*, arXiv:2601.02285.

Tobias Schimanski, Jingwei Ni, Roberto Spacey, Nicola Ranger, and Markus Leippold. 2024a. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17492–17516.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024b. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. DocILE benchmark for document information localization and extraction. In *Document Analysis and Recognition – ICDAR 2023*, volume 14188 of *Lecture Notes in Computer Science*, pages 147–166. Springer.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022a. GriTS: Grid table similarity metric for table structure recognition. *arXiv preprint arXiv:2203.12555*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022b. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4634–4642.

Marijan Soric, Cécile Gracianne, Ioana Manolescu, and Pierre Senellart. 2025. Benchmarking table extraction from heterogeneous scientific documents. *arXiv preprint arXiv:2511.16134*.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *Document Analysis and Recognition – ICDAR 2021*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage DocVQA. *Pattern Recognition*, 144:109834.

Aida Usmanova, Rana Abdullah, Debayan Banerjee, Markus Leippold, and Ricardo Usbeck. 2025. ReportGRI: Automating GRI alignment and report assessment. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4(1):480.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xinyu Wang, Lin Gui, and Yulan He. 2023. A scalable framework for table of contents extraction from complex ESG annual reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13215–13229, Singapore. Association for Computational Linguistics.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. TUTA: Tree-based transformers for generally structured table pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pages 1780–1790.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024 Datasets and Benchmarks Track)*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBERT: A pretrained language model for climate-related text. In *Tackling Climate Change with Machine Learning Workshop, NeurIPS*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, and 1 others. 2026. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11704–11715.

Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Introduces FinTabNet.

A Dataset Details

A.1 `gri_hard` Composition

Table 4 lists the ten `gri_hard` reports underlying the difficulty tiers of §4.

A.2 Page-Label Schemes

Table 7 reports, for each of the 37 PDFs, the declared `/PageLabels` scheme (`trivial`: no entry,

Report	Category	Mod.
HELLENIC_TELECOM_2021	Section-based location	text
OLYMPUS_2021	Section-based location	text
PANASONIC_2022	Section-based location	text
FUJITSU_2023	Complex label scheme	text
ADANI_2022	Double-paged	text
ADAPTEO_2022	Double-paged	text
FUJITSU_2023_x	Scrambled labels	text
HYUNDAI_2020_x	Image-only CI pages	mixed
TOYOTA_2023_x	Image-only	image
TRAIN_187_x	Scrambled labels	text

Table 4: The ten `gri_hard` reports. `_x` suffix = synthetic perturbation of a source report; the rest are real-world adversarial documents. Mod.: *mixed* = only GRI index pages rasterized; *image* = full document rasterized.

Split	PDFs	Pages/PDF median (Q1, Q3)	M	M^{abs}	M^{men}	M^{pg}	$M^{\overline{\text{pg}}}$	$\lvert P_{\mathcal{F}}\rvert/\mathcal{D}$ mean
gri_easy	10	96 (74, 166)	1,936	1,105	831	790	41	5.8
gri_medium	10	128 (92, 148)	1,937	1,007	930	917	13	4.1
gri_hard	10	131 (62, 290)	2,028	1,169	859	823	36	8.2
train_small	7	125 (62, 166)	1,393	895	498	497	1	4.0

Table 5: Per-split statistics. M = total schema rows; M^{abs} / M^{men} = absent / mentioned rows; M^{pg} / $M^{\overline{\text{pg}}}$ = page-indexed / non-page-indexed mentioned rows; $\lvert P_{\mathcal{F}}\rvert/\mathcal{D}$ = mean index pages per report (S1 scale).

implicit $i = l - 1$; `nontrivial`: missing or structurally irregular; `complex`: non-numeric labels requiring custom parsing — see §4 for the tier definitions these feed) and the resulting label-to-index function $i = f(l)$.

A.3 Descriptive Statistics

Table 5 summarises each split along the dimensions that drive the four evaluation stages.

Three patterns stand out. First, absent rows dominate: roughly 55–57% of schema slots are unpopulated across all splits, reflecting real variation in what companies choose to disclose. Second, `gri_hard` has the highest mean index-page count (8.2 vs 4–6 elsewhere), a direct consequence of selecting reports with complex or multi-part content indices. Third, non-page-indexed rows are rare (91 of 3,118 mentioned rows, 2.9%) but must be retained for S2 eval: they carry valid location or status information even when no physical page index can be produced.

Table 6 breaks down rows by disclosure kind. Standard disclosures dominate in both count and mention rate; the topic-qualified family is smaller but carries a distinct difficulty — every mentioned topic-qualified row is page-indexed ($M^{\text{men}}_{\text{top}} = M^{\text{pg}}_{\text{top}}$), so all non-page-indexed rows in the dataset come from standard disclosures. Baselines are evaluated over standard disclosures only.

	Standard			Topic-qualified		
Split	M_{std}	$M^{\text{men}}_{\text{std}}$	$M^{\text{pg}}_{\text{std}}$	M_{top}	$M^{\text{men}}_{\text{top}}$	$M^{\text{pg}}_{\text{top}}$
gri_easy	1,305	774	733	631	57	57
gri_medium	1,305	850	837	632	80	80
gri_hard	1,334	798	762	694	61	61
train_small	928	401	400	465	97	97

Table 6: Rows by disclosure kind. $M_{\text{std}}/M_{\text{top}}$ = standard / topic-qualified total rows; superscripts men / pg = mentioned / page-indexed subsets. $M^{\text{men}}_{\text{top}} = M^{\text{pg}}_{\text{top}}$ in all splits: all topic-qualified rows are page-indexed.

A.4 Qualitative Analysis

The common case — a numeric label resolving cleanly to a physical page — is well represented across all splits (e.g. ℓ_i=`"62"` $\rightarrow$ P_i=$\{61\}$, or a range `"12-15"` $\rightarrow \{11, 12, 13, 14\}$). Three harder cases illustrate why the benchmark is difficult.

Section-based location. In `OLYMPUS_2021` (GRI `102-1`), the content index gives ℓ_i=`"Consolidated Company Outline"` with no page number. The verbatim cell is retained as ℓ_i for auditability, but P_i=$\emptyset$ — the row is mentioned but non-page-indexed. This is the defining challenge for the three section-based reports in `gri_hard`: the location text is present and correct, but no numeric index can be produced without resolving the section pointer.

Topic-qualified disambiguation. In `KERRY GROUP_2023`, the GRI `3-3` row recurs once per material topic in the content index with no topic label — the topic is given only by the surrounding section header. The gold resolves each occurrence to its qualified identifier (e.g. s_i=`3-3:Biodiversity`), yielding thirteen distinct `3-3` disclosures that the flat namespace in `airefinery/esg_cid` collapses to a single row.

Image-only perturbation. In `TOYOTA_2023_x`, all pages are rasterised and carry no text layer. The gold annotation is intact (s_i=`2-1`, ℓ_i=`"3"`, P_i=$\{2\}$), but the disclosure is inaccessible to any text-only system. The same row is trivially extractable from the unperturbed `TOYOTA_2023` in `gri_medium`, making the pair a controlled ablation of what image-only rendering costs.

B Implementation Details

B.1 Evaluation Harness (model-iq)

We run the S2.1 backend sweep (§6.2, §5.3) on **model-iq**, an internal Kubernetes-native control plane for serving and governing both self-hosted and API-based large language models as one fleet.

Report (Company)	Ctry	SASB Industry	Year	Layout	$\lvert P_F \rvert$	G	label scheme	$i = f(l)$
		gri_easy						
		GRI 2016 ($M = 238$)						
TRAIN_170 (Coupa)	US	Software and IT Services	2021	single	6	76	trivial	$i = l - 1$
TRAIN_20 (Pfizer)	US	Biotechnology and Pharmaceuticals	2021	single	10	75	trivial	$i = l - 1$
TRAIN_212 (Molson Coors Beverage)	US	Alcoholic Beverages	2022	single	3	61	trivial	$i = l - 1$
TRAIN_225 (Green Plains)	US	Biofuels	2021	single	10	115	trivial	$i = l - 1$
TRAIN_70 (TJX Companies)	US	Apparel Accessories and Footwear	2022	single	5	76	trivial	$i = l - 1$
		GRI 2021 ($M = 147$)						
AB SKF_2023	SE	Industrial Machinery and Goods	2023	single	5	84	trivial	$i = l - 1$
KERRY GROUP_2023	IE	Processed Foods	2023	single	5	115	trivial	$i = l - 1$
KIA_2024	KR	Automobiles	2024	single	3	102	trivial	$i = l - 1$
KPN_2023	NL	Telecommunications	2023	single	9	63	trivial	$i = l - 1$
TRAIN_90 (SK Inc.)	KR	Asset Management and Custody Activities	2023	single	2	86	trivial	$i = l - 1$
		gri_medium						
		GRI 2016 ($M = 238$)						
TRAIN_143 (Kumbra Iron Ore)	ZA	Metals and Mining	2020	single	5	132	trivial	$i = l + 1$
TRAIN_153 (NTT DOCOMO)	JP	Telecommunications	2020	single	4	148	trivial	$i = l$
TRAIN_187 (Ferrexpo)	CH	Metals and Mining	2020	single	7	84	trivial	$i = l$
TRAIN_194 (Banca Transilvania)	RO	Commercial Banks	2020	single	6	150	nontrivial	$i = l - 1$
TRAIN_307 (Taishin Financial)	TW	Commercial Banks	2021	single	5	114	trivial	$i = l + 1$
		GRI 2021 ($M = 147$)						
HYUNDAI_2023	KR	Automobiles	2023	single	3	108	trivial	$i = l - 1$
HYUNDAI_2024	KR	Automobiles	2024	single	3	108	nontrivial	$i = l - 1$
LACTALIS_2023	FR	Meat, Poultry and Dairy	2023	single	3	57	trivial	$i = l + 1$
TOYOTA_2023	JP	Automobiles	2023	single	3	46	nontrivial	$i = l - 1$
TRELLEBORG_2023	SE	Industrial Machinery and Goods	2023	single	2	83	nontrivial	$i = l + 1$
		gri_hard						
		GRI 2016 ($M = 238/241$)						
ADANI_2022	IN	Engineering and Construction Services	2022	double	8	129	nontrivial	$i = \lfloor l/2 \rfloor$
HELLENIC_2021	GR	Telecommunications	2021	single	14	122	trivial	$i = l - 1$
HYUNDAI_2020_x†	KR	Automobiles	2020	single	5	93	trivial	$i = l - 1$
OLYMPUS_2021	JP	Medical Equipment and Supplies	2021	single	9	148	nontrivial	$i = l$
PANASONIC_2022	JP	Electrical and Electronic Equipment	2022	single	5	137	trivial	$i = l$
TRAIN_187_x†	CH	Metals and Mining	2020	single	7	84	complex	—
		GRI 2021 ($M = 147/148$)						
ADAPTEO_2022	SE	Real Estate Owners, Developers and Investment Trusts	2022	double	3	74	nontrivial	$i = \lfloor l/2 \rfloor$
FUJITSU_2023	JP	Software and IT Services	2023	single	14	117	complex	—
FUJITSU_2023_x† (syn)	JP	Software and IT Services	2023	single	14	117	complex	—
TOYOTA_2023_x† (syn)	JP	Automobiles	2023	single	3	46	trivial	$i = l - 1$

Table 7: Per-report characteristics for all 37 PDFs. $\lvert P_F \rvert$ = number of content-index pages; G = true rows ($\ell_i \neq \varepsilon$); /PageLabels scheme: trivial = no /PageLabels entry; nontrivial = missing or structurally irregular; complex = non-numeric labels requiring custom parsing. $i = f(l)$ is the 0-based page index as a function of the printed label l. † = synthetic perturbation of the preceding real-world report.

This appendix details the parts of its architecture relevant to reproducing the sweep; model-iq's broader scope (cost analytics, RBAC, budget-scoped API keys) is out of scope here.

Unified serving. model-iq deploys open-weight VLMs on GPU-backed infrastructure via **KServe**, with the serving runtime pluggable per model (we use **vLLM** (Kwon et al., 2023) and **SGLang** depending on backend support); GPU scheduling and autoscaling are handled by the GPU Operator and KEDA. API-based closed models (used for the frontier-API references in Table 2) bypass the serving layer entirely and are proxied through the same gateway with provider credential injection. This means every backend — self-hosted or API-based — is reachable through one OpenAI-

compatible request path, so the extraction prompt, image payload, and response parsing code in our S2.1 harness is identical for all 48 backends; only the routed endpoint differs.

Model cards. Each evaluated backend is registered as a *model card*: a versioned record of its identity, weight source (Hugging Face repo id, or provider/model id for closed models), and deployment configuration (serving runtime, GPU/replica count, context-window and generation parameters). Every reported score is tagged with the model-card id and version that produced it, so a score can always be traced back to the exact serving configuration used — important given that quantization, context length, and runtime choice can all shift VLM output quality independently of the underlying checkpoint.

Request path and failure handling. The harness issues one request per confirmed CI page image (§5.3) through the model-iq gateway, which handles auth, rate limiting, and per-backend routing; multi-page CI regions are stitched at the application layer after all page-level responses return. Backends that time out, return malformed output, or exceed context are logged and excluded from that report's aggregate rather than imputed, so per-model N can vary slightly across splits (Appendix A reports per-report coverage).

Why a unified fleet, not per-provider scripts. Prior VLM-vs-OCR benchmarks (§2) typically compare a handful of models, often via each provider's own SDK or a single self-hosted runtime. At 48 backends spanning sub-1B self-hosted VLMs to frontier closed APIs, per-provider integration code would itself become a confound — subtle differences in retry logic, image encoding, or prompt formatting across scripts could shift scores independently of model quality. Routing every backend through one gateway and one model-card schema removes that confound and is what makes the sweep's cross-backend comparisons attributable to the model alone.

Decoding parameters. We do not pin sampling parameters across the sweep: each self-hosted backend reuses its own model family's recommended hyperparameters (temperature, top-p, etc., as specified in the checkpoint's `generation_config.json`), falling back to vLLM/SGLang's defaults (temperature = top-p = 1.0) when a checkpoint specifies none. Closed-model backends use their provider's own default.

B.2 Extraction Prompt and Structured Output

Every S2.1 backend is given the same system prompt and constrained to the same structured-output schema (a Pydantic `LLMExtractedTable` of `ExtractedCIRow` objects), enforced via `response_format` on every request regardless of backend.

System Prompt

```
You are extracting a GRI Content Index table
from a sustainability report page.

For each row extract:
  - standard_id: the GRI standard number,
e.g. "GRI 302-1", "2-1", "102-1"
  - disclosure_text: the disclosure title
or description
  - page_location_text: the page reference
verbatim, e.g. "42", "p.42-43", "N/A"
  - note: any extra remark shown in the row,
e.g. "partially reported", "omitted"

Rules:
  - Include all rows, even those with no
page reference.
  - Do not invent standard IDs - only
extract what is explicitly present.
  - Normalise standard_id to the form
"NNN-N" or "GRI NNN-N" as printed.
  - If the page has multiple index tables
(e.g. side-by-side columns or separate
topic sections), collapse them into one
flat list of rows.
```

Expected Output

```
{
  "rows": [
    {
      "standard_id": "305-1",
      "disclosure_text": "Direct (Scope 1)
GHG Emissions",
      "page_location_text": "44-45",
      "note": ""
    },
    {
      "standard_id": "2-1",
      "disclosure_text":   "Organizational
details",
      "page_location_text": "12",
      "note": "partially reported"
    }
  ]
}
```

B.3 Roster Curation

Table 8 reports the same mean S2-F1 ranking as Table 2 in the main text, with each backend's provider affiliation added as a field; those affiliations line up with the family groupings below. The sweep's

Model	Affiliation	loc-F1	Model	Affiliation	loc-F1
GPT-5.5	OpenAI	0.85	InternVL3.5-4B	Shanghai AI	0.56
GPT-5.4	OpenAI	0.80	InternVL3.5-8B	Shanghai AI	0.56
GPT-5-mini	OpenAI	0.80	Qwen3.5-4B	Alibaba	0.56
Qwen3.5-27B	Alibaba	0.79	**Baseline**	Ours	**0.52**
Gemini-3.5-Flash	Google	0.79	Qianfan-OCR	Baidu	0.52
Gemini-3.1-Flash-Lite	Google	0.78	Step3-VL-10B	StepFun	0.51
Gemini-3-Flash	Google	0.77	InternVL3.5-14B	Shanghai AI	0.50
Qwen3.6-27B	Alibaba	0.77	InternVL3.5-30B-A3B	Shanghai AI	0.50
Ovis2.6-30B-A3B	Alibaba	0.74	Ministral-3-3B	Mistral AI	0.48
Qwen3.6-35B-A3B	Alibaba	0.74	Gemma-4-12B-it	Google	0.45
Gemma-4-31B-it	Google	0.71	Nemotron-3-Nano-Omni	NVIDIA	0.45
Claude Sonnet 4.5	Anthropic	0.71	LLaVA-OneVision-2	LLaVA Team	0.42
GLM-4-6v-flash	Zhipu AI	0.69	UI-TARS-1.5-7B	ByteDance	0.34
Gemma-4-26B-A4B-it	Google	0.67	InternVL3.5-2B	Shanghai AI	0.31
olmOCR-2-7B-1025	AI2	0.66	Gemma-4-E4B-it	Google	0.31
Ministral-3-14B	Mistral AI	0.66	Qwen3.5-2B	Alibaba	0.28
Qwen3.5-35B-A3B	Alibaba	0.65	Granite-Vision-4.1-4B	IBM	0.28
Nemotron-Nano-v2-VL	NVIDIA	0.65	InternVL3.5-1B	Shanghai AI	0.27
Ministral-3-8B	Mistral AI	0.64	MiniCPM-V-4.6	OpenBMB	0.23
Ovis2.5-9B	Alibaba	0.63	Ovis2.5-2B	Alibaba	0.21
Claude Haiku 4.5	Anthropic	0.61	Gemma-4-E2B-it	Google	0.19
GPT-5-nano	OpenAI	0.60	Qwen3.5-0-8B	Alibaba	0.16
Qwen3.5-9B	Alibaba	0.59	Ovis2-1B	Alibaba	0.13
MiniCPM-V-4.5	OpenBMB	0.58	Granite-Vision-3.2-2B	IBM	0.06

Table 8: Mean loc-F1 across `gri_easy`, `gri_med`, and `gri_hard` (oracle S1). Closed-source models are indicated in blue; the Baseline (§5.1) is indicated in red.

48 backends are drawn from the following model families:

- **Claude**: claude-haiku-4-5 and claude-sonnet-4-5 (Anthropic, 2025a,b)
- **Gemini**: gemini-3-flash-preview (Google, 2025), gemini-3.1-flash-lite (Google DeepMind, 2026a), and gemini-3.5-flash (Google DeepMind, 2026b)
- **Gemma**: the gemma-4 family (gemma-4-e2b-it, gemma-4-e4b-it, gemma-4-12b-it, gemma-4-26b-a4b-it, and gemma-4-31b-it) (Gemma Team, 2026)
- **GLM**: glm-4-6v-flash (Hong et al., 2025)
- **GPT**: gpt-5-mini and gpt-5-nano (Singh et al., 2025), gpt-5-4-1 (OpenAI, 2026a), and gpt-5-5 (OpenAI, 2026b)
- **Granite Vision**: granite-vision-3-2-2b (Granite Vision Team et al., 2025) and granite-vision-4-1-4b (IBM Granite Vision Team, 2026)
- **InternVL**: the internvl3-5 family (internvl3-5-1b, internvl3-5-2b, internvl3-5-4b, internvl3-5-8b, internvl3-5-14b, and internvl3-5-30b-a3b) (Wang et al., 2025)
- **LLaVA**: llava-onevision-2-8b-instruct (An et al., 2026)
- **MiniCPM-V**: minicpm-v-4-5 (Yu et al., 2026) and minicpm-v-4-6 (Yu et al., 2026; OpenBMB, 2026)
- **Ministral**: the ministral-3 family (ministral-3-3b, ministral-3-8b, and ministral-3-14b) (Liu et al., 2026)
- **Nemotron**: nemotron-nano-v2-vl (Deshmukh et al., 2025) and nemotron-3-nano-omni-reasoning (Deshmukh et al., 2026)
- **olmOCR**: olmocr-2-7b-1025 (Poznanski et al., 2025)
- **Ovis**: ovis2-1b (Lu et al., 2024), the ovis2-5 models (Lu et al., 2025), and ovis2-6-30b-a3b (AIDC-AI Ovis Team, 2026; Lu et al., 2025)
- **Qianfan**: qianfan-ocr (Dong et al., 2026)
- **Qwen**: the qwen3-5 family (qwen3-5-0-8b, qwen3-5-2b, qwen3-5-4b, qwen3-5-9b, qwen3-5-27b, and qwen3-5-35b-a3b) (Qwen Team, 2026a) and the qwen3-6 models (qwen3-6-27b and qwen3-6-35b-a3b-2) (Qwen Team, 2026b,c)
- **Step3**: step3-vl-10b (Huang et al., 2026)
- **UI-TARS**: ui-tars-1-5-7b (Qin et al., 2025)

Table 8 reports the same mean S2-F1 ranking as Table 2 in a two-column layout with affiliation as an added field. We print it in §B.3 rather than repeating it here, since the affiliation column pairs directly with the family-by-family roster listed there.

Evidence-Grounded Auditing of Identification Assumptions in Climate-Policy Causal Evaluations

Yonghong Zhang[1] **Yong Xie**[2] **Isabel M. Parra**[1] **Ricardo Correia**[1]
[1]Department of Finance, Universidad Autónoma de Madrid, Madrid, Spain
[2]Spanish National Research Council (CSIC), Madrid, Spain
Correspondence: yonghong.zhang@estudiante.uam.es

Abstract

Difference-in-differences (DID) studies are widely used to evaluate climate policy, but assessing the evidence supporting their identification assumptions remains challenging. We introduce ARGUS, a structured language-model pipeline that audits reported evidence against an eleven-dimension assumption–implication–evidence rubric and abstains when relevant evidence cannot be retrieved. We evaluate ARGUS using injected flaws, economics papers, and a small pilot with reconciled labels. On the 11-flaw benchmark, ARGUS detects 73% of planted flaws, compared with 18% for a keyword-based pipeline. Across 26 economics papers, ARGUS abstains on about 40% of paper–dimension assessments for lack of retrievable evidence. In a five-paper pilot with labels reconciled by two annotators, it assigns a higher risk level than the labels on 25 of the 33 assessments it completes. A rule fixed before the labels arrived removes most of this in-sample; weighted agreement stays low. ARGUS provides evidence-linked risk reports that localize potential weaknesses for expert review, without adjudicating causal claims. Code and data: https://github.com/yonghongzhang-io/ARGUS.

1 Introduction

Difference-in-differences (DID) is widely used to evaluate climate and environmental policy, including carbon pricing, emissions-trading schemes, and emission standards. The credibility of these estimates depends on identification assumptions such as parallel trends, no anticipation, appropriate treatment of staggered adoption, and limited interference between units, each of which a large body of methodological work has shown to be fragile (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfœuille, 2020; Roth et al., 2023). These assumptions can be hard to assess in climate settings, where policies may overlap, be phased in over time and spill across administrative boundaries, and outcomes may be modelled or remotely sensed (Appendix A). Where such estimates inform policy, causal conclusions built on weakly supported assumptions can mislead.

Assessing an identifying assumption requires more than locating a reported diagnostic. A paper may present parallel trends over an event-study figure, leaving the reader to weigh the design, the pre-period estimates, the inference, and the policy timing before accepting the claim; the assessment has to be repeated for every assumption and every paper. This is one facet of a broader concern with the reliability of empirical findings (Ioannidis, 2005; Silberzahn et al., 2018; Brodeur et al., 2020).

We argue that identification credibility is auditable even when the true causal effect is not, building on the research-design tradition in applied econometrics (Angrist and Pischke, 2010; Athey and Imbens, 2017; Rambachan and Roth, 2023). Because the counterfactual is unobserved, ARGUS (named after Argus Panoptes, the hundred-eyed watchman) does not ask whether an estimated effect is true. Instead, it asks whether each identifying assumption is adequately supported by the evidence reported in the paper, and where that support is weakest. ARGUS represents a DID study through an 11-dimension *assumption–implication–evidence* rubric and uses a bounded, retrieval-gated pipeline built on a large language model (LLM) to retrieve relevant evidence, assess its adequacy, and produce an evidence-grounded risk report for expert review (Figure 1). Model agency is confined to scoped retrieval and assessment tasks behind a fixed, deterministic control flow, unlike free-roaming agents whose control flow is model-decided (Yao et al.,

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 199–219
October 28, 2026 ©2026 Association for Computational Linguistics

2023); ARGUS flags potential weaknesses rather than adjudicating causal truth.

Evaluating such an auditor is difficult because the true policy effect is unobserved. We therefore use *flaw injection* as local ground truth: starting from a supported study fixture, we introduce one known identification flaw and test whether ARGUS detects it, remains quiet on the clean version, and localizes the affected dimension. We complement this controlled evaluation with a run over 26 corpus papers tagged DID and a small diagnostic pilot with reconciled labels, used to identify calibration and applicability failures. Climate-policy evaluation motivates the rubric; the present evaluation uses synthetic environmental-policy fixtures and a general-economics corpus, so performance on a dedicated climate-policy corpus is untested.

Contributions. We develop and evaluate a rubric and pipeline for assessing the reported evidence behind DID identification assumptions, without access to the causal effect:

1. an *operationalization*: a machine-auditable *assumption–implication–evidence* rubric and a bounded, evidence-grounded pipeline with explicit abstention (§3);
2. an *evaluation methodology*: planted identification errors as local ground truth, with compute-graded and oracle ablations that locate the gain in per-dimension prompting and trace every synthetic omission miss to a gate abstention (§5);
3. a *diagnosis*: on real papers, grounding limits *coverage* while over-severity limits *agreement*, and a pre-specified rule derived on the same papers reduces the latter in-sample.

2 Related Work

DID identification and its threats. A large econometrics literature formalizes when DID estimators fail under staggered timing and heterogeneous effects: two-way fixed effects' use of already-treated units as controls (Goodman-Bacon, 2021), negative weights (de Chaisemartin and D'Haultfœuille, 2020), robust estimators (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021), assumption and pre-trend surveys (Roth et al., 2023), and fragile inference under serial correlation (Bertrand et al., 2004). ARGUS adds no estimator: it *operationalizes* this literature as an auditable rubric, turning each known threat into a dimension whose supporting evidence can be checked in a given paper (§3).

LLM causal reasoning and benchmarks. A parallel line asks whether LLMs reason about causality. Benchmarks (CLadder (Jin et al., 2023), Corr2Cause (Jin et al., 2024), CausalBN-Bench (Zhou et al., 2026)) and analyses (Kıcıman et al., 2024; Yang et al., 2024; Zečević et al., 2023) report mixed evidence: strong performance on some causal tasks, but models often reproduce causal *language* without reliable causal *inference*; a few move toward applied settings (Lee et al., 2025) and toward separating *identification* from *estimation* (Sawarni et al., 2026). ARGUS addresses a different task: whether a *human* paper's identification assumptions are adequately evidenced. A causal verdict is out of scope by design.

Responsible AI for empirical research. Research reliability is a long-standing concern, documented by wide analyst-driven result spreads (Silberzahn et al., 2018; Botvinik-Nezer et al., 2020) and by specification searching in causal economics (Brodeur et al., 2020), and artificial intelligence (AI) is increasingly proposed as a tool for it. LLMs are also used directly as evaluators and reviewers, judging model outputs (Zheng et al., 2023), giving feedback on research papers (Liang et al., 2024), and reviewing inside automated-science pipelines (Lu et al., 2024); these evaluators already produce ratings, pairwise preferences, and structured reviews. ARGUS applies an assumption-specific rubric to DID studies, links each dimension to in-paper evidence, and is measured against injected ground truth. ARGUS focuses on the evidence for *identification credibility* specifically (the assumptions supporting the causal interpretation) and issues no verdicts, producing evidence-grounded, human-reviewable reports that localize risk for the expert's final call. Closest to our setting, CausalVerify (Zhang et al., 2026a) grounds evaluation in *execution*: it runs a model's generated estimation code against synthetic data and checks the recovered effect. That benchmark does not audit identification assumptions, which is the layer ARGUS takes up; we reuse its released corpus of economics papers and add a dimension-level identification-risk annotation layer on top of it.

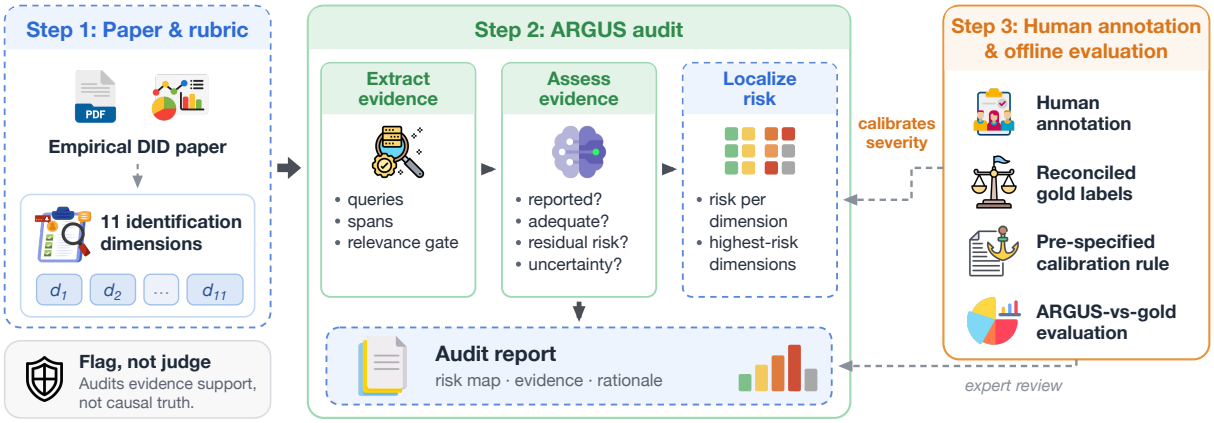


Figure 1: Overview of ARGUS. Step 1 converts an empirical DID paper into an auditable structure using an 11-dimension identification rubric. Step 2 retrieves dimension-specific evidence, assesses reported-evidence adequacy with a bounded LLM audit module, and localizes risk in a human-reviewable report. Step 3 uses labels from two annotators, reconciled by them, to evaluate audit severity and score a pre-specified calibration rule offline; deployment requires only expert review of the report. Colours as in Figure 2: blue, deterministic steps; green, model calls; orange, human steps.

Natural language processing for climate evidence. Natural language processing (NLP) systems for climate text verify claims against evidence (CLIMATE-FEVER (Diggelmann et al., 2020)), detect environmental claims (Stammbach et al., 2023), analyse corporate climate disclosures with traceable LLM answers (Ni et al., 2023), and tune models for faithful evidence-based question answering (Schimanski et al., 2024). The verification and question-answering systems ground answers in documents; ARGUS extends document-grounded verification from *claims* to the *identification argument* of a causal study, auditing whether reported evidence supports each assumption.

Evaluation methodology and bounded agents. Our evaluation follows benchmarking lessons: holistic evaluation (Liang et al., 2023) over a single number; dynamic, adversarial benchmarking (Kiela et al., 2021) (which inspires our controlled flaw injection); and the caution that operationalizing a capability is non-neutral (Raji et al., 2021; Jacobs and Wallach, 2021) (§6). A closely related concern is that an agent benchmark can measure its harness rather than its model: Zhang et al. (2026b) show that a fixed scaffold making the execution-critical decisions and a scorer grading output shape mask each other, so repairing either alone leaves the ranking uninformative. Our ablations apply a related check inward: they compare joint and per-dimension prompting, with and without retrieval and relevance gating, holding the model and rubric fixed (§5.2). Where ReAct-style agents (Yao et al., 2023) choose their own action sequence, ARGUS fixes the order of retrieval and assessment and confines model calls to those two stages, trading autonomy for inspectability and evidence-traceability.

3 The ARGUS Framework

3.1 Identification rubric

ARGUS decomposes a DID study into eleven auditable dimensions of identification, design, inference, and reporting: parallel trends, no anticipation, staggered-timing handling, SUTVA (stable unit treatment value assumption, no interference) and spillovers, control-group construction, specification, inference, sample-period selection, concurrent policies, placebo/robustness, and data measurement. Each dimension is expressed as an *assumption* or design requirement (what must hold), a testable *implication* (the diagnostic that should then be observable), and the *evidence* a credible paper would report to support it. For a carbon-pricing or emissions-trading evaluation, for instance, the parallel-trends dimension expects an event-study or pre-trend test with pre-period estimates near zero (an insignificant pre-trend alone can reflect low power; Roth, 2022), while no-anticipation turns on the gap between a scheme's *announcement* and *implementation* dates. The rubric is declarative, stored in a configuration file (`identification_dimensions.yaml`), so the audited dimensions are explicit and in-

spectable rather than buried in code. A companion flaw taxonomy (§4) maps each known identification threat one-to-one onto a dimension, which both documents the rubric's coverage and supplies the perturbations used for evaluation.

3.2 Bounded LLM pipeline

The audit runs a fixed, deterministic sequence, *decomposition* → *extraction* → *assessment* → *localization* → *report*, shown in the top band of Figure 2. The stage order is written in code and never chosen by a model. *Decomposition* instantiates the rubric for the paper at hand; *extraction* gathers, per dimension, the in-paper evidence relevant to that assumption; *assessment* evaluates each assumption–implication–evidence chain and emits a per-dimension risk judgement with a rationale and cited evidence; *localization* aggregates these judgements into a risk map that ranks the dimensions assigned the highest risk (ties broken by the fixed rubric order); and *report* renders a transparent, evidence-grounded document for a human reviewer.

Model calls are confined to two stages, extraction and assessment. In the configuration evaluated in §5 they resolve, for each dimension, to a fixed sequence: deterministic lexical retrieval over the paper's sections and captions, one relevance-gate call that labels each retrieved passage *high*, *partial*, or *none*, and, if at least one passage is kept, one adequacy call. If no passage is kept the dimension is scored `unknown` without an adequacy call; a dimension with only partially relevant passages is still judged (Appendix D). Figures and tables enter only through their captions. The code also defines a more general bounded loop for these two stages (a fixed tool set of `evidence_search`, `figure_parse`, and `policy_lookup`; a hard step budget, `max_steps`; a logged trace of every step and cited passage); the reported experiments run it with `max_steps=1` and never invoke `figure_parse` or `policy_lookup`, and all other stages are plain orchestration with no model call. We therefore call the design *bounded* rather than agentic: the evaluated system is structured, dimension-scoped prompting over retrieved evidence with an explicit abstention path, not an open-ended agent.

Why bound the model's role. Fixing the stage order makes the sequence of operations explicit and comparable across runs and keeps every judgement anchored to logged evidence a human can inspect. It does not make model outputs identical: Appendix F reports the variation observed across repeated runs. The expert sits at the end of the pipeline: ARGUS surfaces and localizes identification risk; the judgement itself stays with the human.

3.3 Output

ARGUS emits study–dimension–risk triplets, each with an evidence-grounded rationale, alongside the human-readable report. In the evaluated prompts (Appendix E) *low*, *medium*, and *high* correspond approximately to retrieved evidence judged sufficient, partial, and missing or flawed; `unknown` is produced by the gate when no relevant passage is retained and reflects retrieval, not a judgement about the paper. None of the four is a verdict on the causal effect. This output shape keeps the system both auditable by a human and measurable by the evaluation that follows.

4 Evaluation by Flaw Injection

4.1 Local ground truth without the counterfactual

The defining difficulty in evaluating a causal-identification auditor is that the quantity of ultimate interest (the true causal effect of a carbon price, an emissions cap, or an emission standard) is never observed. We therefore do not attempt to certify that ARGUS recovers true effects. Instead we construct *local* ground truth by controlled perturbation. Beginning from a short synthetic fixture, an environmental-policy DID study that reports supporting evidence on all eleven dimensions, we inject exactly one known flaw that targets a single dimension, drawn from the flaw taxonomy (`flaw_taxonomy.yaml`; Appendix B traces each flaw to its methodological basis). The target dimension is known by construction, and the expected response (risk of at least *medium* on that dimension) is defined by the rubric, without ever needing the true effect. The bottom band of Figure 2 shows the loop: the clean and injected versions are each audited, and three metrics are computed.

- **Detection**: does the injected version raise risk on the targeted dimension to *medium* or above? (All reported results use this threshold.)

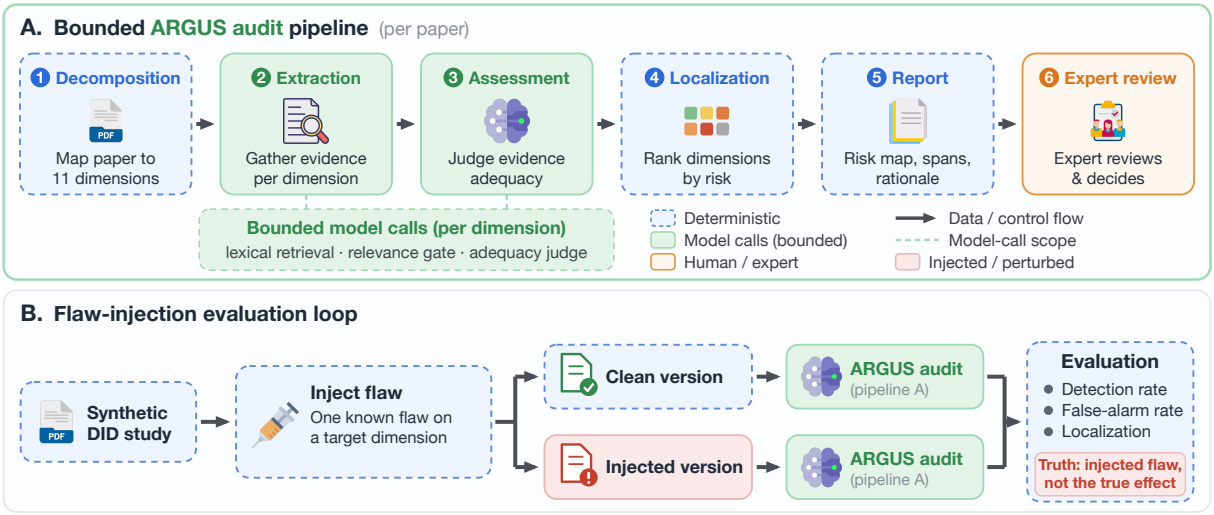


Figure 2: ARGUS architecture. The top band is the audit pipeline, a fixed deterministic sequence; model calls are confined to two stages (extraction and assessment). As evaluated, these resolve per dimension to a fixed call sequence (lexical retrieval, one relevance-gate call, one adequacy call) with a step budget of one and a logged trace; the designed tools `figure_parse` and `policy_lookup` are not exercised (§3.2). The bottom band shows the flaw-injection evaluation loop, which injects one known flaw, audits the clean and injected versions, and scores detection, false alarm, and localization against the injected flaw as local ground truth, never against the unobserved true effect.

- **False alarm**: does the clean version stay quiet on that dimension?
- **Localization**: is the targeted dimension the one the auditor flags most strongly?

4.2 Structural perturbation, not sentinels

How a flaw is injected is a methodological choice that matters. A naïve injector appends a sentence that *names* the flaw: e.g. "treated and control have non-zero pre-period coefficients with significant leads." If the detector then searches for exactly those phrases, detection becomes a tautology: the injector plants the very string the detector matches on. In our initial implementation this produced a detection rate of 1.000 inflated by phrase leakage.

We replace this with *structural perturbation* of two kinds. *Omission* flaws *remove* the section, figure, or table that supplies the evidence the rubric specifies for a dimension (e.g. deleting the placebo-robustness section or the balance table). They test sensitivity to missing reported support; they do not establish that the underlying assumption is violated. *Commission* flaws *rewrite* a section to introduce a stated design problem in prose intended to resemble how a flawed paper would describe it (e.g. pooling all province-years in a two-way fixed-effects regression under staggered policy timing), *without* the phrases the keyword detector looks for. A regression test fails if an injected paper contains the target dimension's predefined negative-signal phrases, so this phrase leakage cannot silently return; detection cannot come from matching those phrases.

5 Results

We evaluate ARGUS along four axes: a synthetic flaw-injection protocol, an architecture ablation, a run over 26 papers tagged DID in an economics corpus, which stress-tests the audit architecture across empirical economics beyond the climate-policy setting that defines its target deployment, and a pilot with reconciled labels. Four findings emerge: (i) LLM pipelines that judge evidence *adequacy* detect planted flaws that a keyword-presence pipeline misses; (ii) per-dimension prompting detects more planted flaws than a single joint call, while the retrieval gate turns retrieval failure into an auditable `unknown` signal; (iii) on real papers retrieval limits coverage, and weakly relevant passages are associated with high-risk ratings; and (iv) against the reconciled labels ARGUS is systematically over-severe, a bias that a pre-specified rule derived on the same papers reduces in-sample.

Table 1: Synthetic flaw injection on the `clean_supported` fixture, eleven flaws; all LLM rows gpt-4o, detection replicated on the dated 2024-11-20 snapshot; see Appendix F on run-to-run differences. Every row is an end-to-end pipeline with its own evidence path (keyword and two-stage do not share retrieved evidence, Appendix D; Appendix K holds it fixed); the ablation arms share the model and rubric but remove retrieval, the gate, or both. The final row is the leaked, pre-fix circular setup. On the 33-variant benchmark the two-stage assessor rates one clean dimension medium, a 0.09 over-flag (§5.2). Wilson 95% intervals for the main rates are in Table 8.

assessor	calls	detection	false alarm	localization
keyword baseline	0	0.182	0.000	0.182
single-pass	1	0.818	0.000	0.818
per-dimension, no retrieval	11	1.000	0.000	1.000
two-stage ARGUS	22	**0.727**	**0.000**	**0.727**
(circular, pre-fix)	–	1.000	0.000	1.000

5.1 Adequacy pipelines versus keyword presence; what the gate trades

We compare three assessors under one flaw-injection protocol on the `clean_supported` fixture (one flaw injected per run, all eleven flaws; `max_steps`=1). A deterministic **keyword baseline** scores each dimension from the presence of supporting and risk-indicating vocabulary. The **two-stage ARGUS assessor** judges evidence *adequacy* (whether the reported evidence supports the assumption under the rubric). The two are complete pipelines with their own evidence paths: the baseline scores paragraph chunks, while the two-stage path re-retrieves and gates whole sections (Appendix D). The pair is therefore an *end-to-end* comparison and cannot separate the judgement policy from retrieval. A **single-pass** ablation removes the architecture entirely: the full paper and all eleven dimensions in one `gpt-4o` call, with the same model, temperature (the sampling-randomness setting, held at 0), risk schema, and scoring harness, but no staged retrieval, relevance gating, or bounded tool use; its strict output schema admits only low, medium, or high, so this arm cannot abstain (Appendix E). Table 1 reports the headline metrics.

The keyword baseline is blind to commission flaws. The baseline never raises a false alarm and correctly flags the two omission flaws (`measurement_break` and `spillover_contamination`) by the *absence* of their vocabulary, but beyond those detection is only 0.182 (Figure 3b). Two failure modes explain it: the same keywords occur in other sections (deleting the placebo section does not help when another section also mentions "placebo") and blindness to *every* commission flaw, whose flawed-but-plausible evidence carries the expected topic vocabulary.

Detection improves with the LLM pipelines. Replacing the keyword pipeline with the two-stage LLM pipeline raises detection from 0.182 to 0.727 with no false alarms, catching *every* commission flaw the keyword scorer is blind to; handing each dimension the full fixture (the per-dimension arm) reaches 1.000. Since retrieval also differs, a pre-specified control holds the evidence fixed (Appendix K). Reading the keyword pipeline's own evidence chunks, with no gate, the adequacy judge detects 10 of 11 flaws against the keyword scorer's 2 (exact McNemar p=0.008), and 31 against 8 of the 33 variants. The gain is a trade: the judge rates one clean dimension medium and 11–16% of non-target dimensions on injected papers, where the keyword scorer flags none. The model reads flawed-but-natural prose the keyword scorer cannot; the audit report in Appendix I shows it flagging the injected anticipation flaw with the announcement-gap rationale, with no sentinel keyword present.

What the gate trades. The gated pipeline detects *less* than its own ablations here (0.727 vs 0.818 and 1.000) by design: its three misses are all *omission* flaws on which retrieval finds nothing and it abstains rather than guessing; in the reported runs each of these misses terminates at the relevance gate as `unknown`, and the oracle-retrieval arm of §5.2 flags the same cases when the assessor is called on the target section. The gate's effect is behavioural and shows on real papers. A single-pass variant whose prompt *does* permit `unknown` ("use sparingly"; Appendix E) abstained on one of 275 cells (0.4%), against 38.9% mechanical abstentions for the gated pipeline (Appendix C). The designs behave differently when evidence is hard to ground; without cell-level gold the contrast does not show which is more reliable.

Scope. These are short synthetic fixtures; ceilings on clearly injected flaws are a proof-of-concept, tested next on 33 variants (§5.2) and real papers (§5.3).

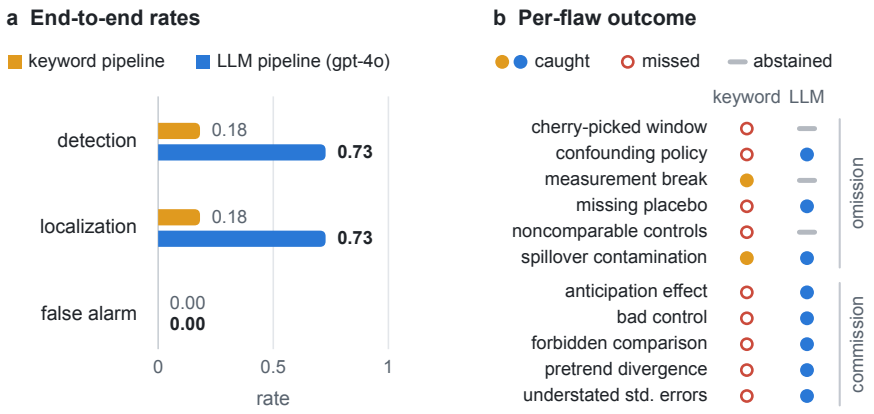


Figure 3: Keyword pipeline vs LLM evidence-adequacy pipeline (n=11 flaws, structural injection). **(a)** End to end, the LLM pipeline raises detection and localization $0.182 \rightarrow 0.727$ with false alarm 0 (the two pipelines retrieve evidence differently; Appendix K holds it fixed). **(b)** Per-flaw caught/missed: the keyword baseline is *blind to every commission-type flaw*, all of which the LLM catches; the LLM's three misses are omissions on which it abstains. Table 2 reports the 33-variant benchmark.

5.2 An expanded benchmark of 33 flaw variants

The eleven flaws are deliberately clear-cut, so we built an expanded benchmark: 33 variants, three per dimension (two *commission* rewrites, one *omission*). The rewrites were written to resemble descriptions of flawed design choices and are verified free of the keyword detector's predefined negative-signal phrases. Three two-stage runs (gpt-4o, temperature 0) are near-deterministic (detection identical on 32/33 variants, risk labels on 31/33; Appendix F). Detection is 0.73–0.76 and localization 0.64–0.70, in line with its behaviour on the eleven clear flaws.

Where the gain comes from. A compute-graded ablation on this benchmark (Appendix C) removes retrieval and gating in turn, the descaffolding check of Zhang et al. (2026b) applied to our own pipeline: a full-paper single pass (one call) detects 0.727; per-dimension prompting without retrieval (eleven calls) detects 0.879; the full two-stage pipeline (twenty-two calls) detects 0.747. Per-dimension prompting detects more variants than a single joint call (29/33 vs 24/33); the comparison changes both the task decomposition and the number of calls, so it does not isolate their separate contributions. The retrieval-and-gate layer then *trades* omission detection for abstention (on these short fixtures a gateless assessor can tell from its context that the section is missing; on real papers it commits judgements nearly everywhere, §5.3). An *oracle-retrieval* arm supplies a diagnostic comparison: fed the target section directly (empty for omissions, bypassing the gate), the assessor detects 0.86 of commissions and 1.00 of omissions, while the gate suppresses the alarms section-level evidence alone triggers (0.21 oracle arm vs 0.09; the full-context arm's 0.00 reflects its wider view). At n=33 none of these pairwise arm differences reaches significance under an exact paired McNemar test (single-pass vs per-dimension p=0.13, per-dimension vs two-stage p=0.22, two-stage vs oracle p=0.13; Appendix G); the claim concerns the observed execution path (every extra omission the oracle arm catches is a case the gated pipeline had scored `unknown` at the gate), not a significant rate gap.

Cross-model robustness. Porting the unchanged pipeline across providers shows that high commission detection is not model-specific (Appendix C): *commission* detection is high everywhere (gpt-4o 0.89, Claude Opus 4.8 1.00, Gemini 2.5 Flash 0.91, Llama 3.1 8B 0.95), but the clean-fixture flag rates differ substantially (Table 5), so high detection alone does not show that each model separates the flawed and clean versions reliably. What differs is the operating point, and Table 6 shows where: Opus 4.8 rates the *clean* fixture medium or high on 18 of 22 commission pairs (false alarm 0.82), the over-severity the human gold finds for gpt-4o, only stronger; Llama 8B almost never abstains (one `unknown` in eleven omissions) and calls the missing evidence high; Gemini is closest to gpt-4o (26/33 identical injected verdicts). High commission detection is observed with every tested model; the abstention

Table 2: The 33-variant benchmark (gpt-4o, mean of three runs at temperature 0; detection decisions identical on 32/33 variants). Adequacy reasoning catches *commission* flaws; *omission* misses are abstentions to `unknown`, marking the evidence-grounding bottleneck rather than wrong judgements. Intervals in Table 8.

variant type	n	detection	localization
commission	22	0.89	0.82
omission	11	0.45	0.33
all	33	**0.75**	**0.66**

behaviour and the false-alarm profile are not, so changing the model requires a fresh assessment of calibration and false alarms (§5.6).

Detection differs by flaw type (Table 2). The assessor catches 0.86–0.91 of *commission* flaws (flawed-but-present evidence, the case adequacy reasoning targets) but only 0.45 of *omission* flaws. The omission misses are not confident errors: the six missed variants all abstain to `unknown` rather than assert low risk, because deleting the supporting section makes retrieval fail. This is the same failure point we find on real papers (§5.3): on these fixtures the assessor flags most commission variants when the evidence is retrieved and abstains when it is not. The only false-alarm signal (0.09 in every run) is one clean-fixture dimension rated medium (*inference* in run 1, *specification* in runs 2 and 3), the same over-severity we quantify against the human gold (§5.5); the clean audit is itself not deterministic. Table 1 shows no false alarm on the original eleven-flaw set.

5.3 Real papers: retrieval limits audit coverage

We next audit the 26 papers tagged difference-in-differences in the CausalVerify corpus (Zhang et al., 2026a) (Appendix H gives its composition and the limits of that tag). There is no injected ground truth; we report the raw per-dimension risk distribution of each assessor (Figure 5). The keyword baseline is permissive: 91% of its judgements are *low*. It sees topic vocabulary, declares the dimension supported, and flags almost nothing. The two-stage LLM assigns a wider range of risk labels, and its most informative output is that ∼40% of judgements are `unknown`: the relevance gate could not surface evidence, so the system says so rather than over-flagging. Two regimes appear: dimensions whose discussion is diffuse or lives in figures the text retrieval cannot reach are dominated by `unknown`, while dimensions where a section is retrieved are judged inadequate. On real papers, then, retrieval limits coverage, and weakly relevant passages are associated with high-risk ratings. These results identify retrieval and the treatment of missing evidence as failure points; they do not separate their contributions from rubric applicability and assessment errors. The design therefore reports retrieval quality and an explicit `unknown` state distinct from a substantive high-risk judgement.

Why fixtures abstain but real papers flag. The synthetic omission misses of §5.2 become abstentions, yet on real papers a missing check usually becomes a substantive *high*. The asymmetry is one branch of the gate. Retrieval is lexical (Appendix D): deleting a section from the short fixture leaves no passage that even shares its vocabulary, the gate finds nothing relevant, and the pipeline abstains without calling the judge. A full paper usually contains a section that shares the vocabulary, so the gate returns *partial* relevance, the judge is called on an off-target passage, reports the evidence as missing, and the prompt maps missing to high. The pilot shows the pattern (Table 7): all 22 failed-retrieval cells are `unknown`, and 20 of the 22 weak-retrieval cells are *high*, 19 of them with `evidence_status=missing`. This is the branch the calibration layer of §5.6 targets.

Judgements cite specific evidence (Appendix I gives a qualitative example). Roughly 44% of judgements on these papers are *high*; without dimension-level human labels beyond the five pilot papers we cannot say whether this is warranted or over-strict where retrieval succeeds.

5.4 An annotation pilot with reconciled labels

To test agreement with human judgement, we built a small reconciled reference set (the "gold"): two annotators from outside the author team, who had no part in building ARGUS, independently labelled five corpus papers × 11 dimensions (55 cells), without AI tools or sight of ARGUS's output, and then reconciled the 13 cells on which they differed in applicability or risk between themselves, without the first author (Appendix J). Their independent labels agree on 45 of the 54 cells both rated (Cohen's $\kappa = 0.56$; quadratic-weighted $\kappa = 0.62$), every disagreement is one severity step, and 23 of the 55 cells are labelled under an *analogue* reading of the dimension because the de-

sign is not a canonical DID. The reconciled gold (55 cells) is low 7, medium 46, high 2; we read this pilot as preliminary. The reviewed version used labels two people had produced with LLM assistance and sent to the first author, with no annotation record; they are replaced here (agreement with the human gold 31/55, Appendix J).

5.5 ARGUS versus the human gold

We use this gold as a *diagnostic pilot*, not a validation of accuracy: it localizes *how* ARGUS fails (Table 13, Figure 4). ARGUS abstains on 22/55 cells, and when it answers its errors are almost entirely one-directional: 25/33 answered cells are more severe than the reconciled gold, none less severe; the gold contains two high-risk cells, whereas ARGUS assigns 24. With so few high gold labels, precision and recall are noisy (of the two high gold cells ARGUS flags one and abstains on the other). Most abstentions are retrieval failures rather than absent evidence: both annotators located reported evidence on 19 of the 22 abstained cells, at least one on 21.

5.6 Calibration: a pre-specified rule, in-sample

Of the re-run's 24 over-severe *high* judgements, $23/24$ have `evidence_status=missing` and $20/24$ have `retrieval_quality=weak`. ARGUS treats "weak retrieval found nothing" as a substantive *high*, conflating a *retrieval failure* with the paper lacking the evidence, which the reconciled human label puts at low or medium: a fixable calibration target. The pattern re-emerges in every leave-one-paper-out fold of the pilot (a recurrence check; no rule is refitted). (This uses a re-run with retrieval-quality and evidence-status fields; it differs from the corpus run of §5.5 in one cell; its pre-calibration figures are the same 25/33 over-severe and exact agreement 0.24.)

A deterministic rule over fields the LLM already produces reduces the over-severity: a weak-retrieval *high* (20 cells) is demoted to *medium* (keep coverage) or to `unknown` (abstain). Keeping coverage, the rule raises exact agreement from 0.24 to 0.76 ($8/33$ to $25/33$) and cuts over-severe cells from 25 to 8: on the 33 answered cells it changes seventeen incorrect labels to the reconciled label and no correct label to an incorrect one, and keeps the one high gold cell ARGUS had caught. The abstention variant removes those 20 predictions (coverage $33/55 \rightarrow 13/55$) and leaves the remaining labels unchanged. Three further single-cell rules lift exact agreement to 0.85 without raising weighted κ under the demotion policy; they are post-hoc string and field rules, reported as exploratory (Appendix J, Table 14, Figure 6). Exact agreement is a weak yardstick here: the gold is 84% *medium*, and a constant *medium* label would score $26/33$ on the same cells; the statistic to read is weighted κ, $0.08 \rightarrow 0.23$, with an interval that includes zero (Appendix G). The rules were fixed before the human labels arrived but were derived on these same five papers against the earlier labels, so the lift is in-sample: it points the same way in all five papers and is untested on unseen papers.

6 Discussion

What ARGUS measures. A measurement is only as good as the link between the construct and the task that operationalizes it (Jacobs and Wallach, 2021; Raji et al., 2021). ARGUS audits whether a climate-policy study's identification assumptions are *adequately supported by its reported evidence*: reported-evidence adequacy is an operational proxy for identification credibility, not causal validity itself.

Why climate policy. Climate-policy estimates inform governance, so a confidently wrong credibility judgement has a cost; ARGUS therefore abstains when its gate finds no relevant passage.

Retrieval limits audit coverage. A further gap is retrieving the relevant evidence. On the real-paper corpus it limits coverage, and in the pilot an assessor handed weakly relevant evidence usually rates the dimension high (20 of 22 weak-retrieval pilot cells), so ARGUS reports retrieval quality and an explicit `unknown` distinct from a substantive high-risk judgement.

7 Conclusion

The causal effect of a climate policy is unobservable; whether a study's reported evidence supports its identification assumptions can be assessed. ARGUS operationalizes that assessment with a bounded, evidence-grounded pipeline, flaw injection as local ground truth, and a reconciled pilot whose human labels evaluate the auditor and score a pre-specified rule that reduces its over-severity in-sample; its value in expert review and on climate-policy papers is untested.

Limitations

Our human gold is pilot-scale: 5 papers and 55 cells from two annotators who reconciled their differences themselves, so the comparison is a preliminary signal rather than a calibrated benchmark; its labels are 84% *medium*, which makes exact agreement lenient and κ the more informative statistic. The gold is the reconciled judgement of two annotators, not of a panel of DID methodologists; it has only two *high* cells; 23 of its 55 cells judge an analogue of the assumption because the design is not a canonical DID; the main ARGUS-versus-gold figures are conditional on the 33 cells ARGUS answered (it abstained on 22, on 19 of which both annotators found evidence); and a constant *medium* label would out-score the calibrated system on exact agreement. The current scope is confined to difference-in-differences identification; whether the rubric-plus-injection methodology transfers to other quasi-experimental designs is untested. The rubric, flaw taxonomy, and injector are co-designed, so synthetic detection rates are best read as internal consistency under known threats rather than external generalization (Appendix B traces each injected flaw to its methodological basis, so the threats are inherited while their operationalization and the injector's prose are ours); flaws authored by independent experts are future work. The benchmarks are small: at n=11 and n=33 no pairwise difference between architecture arms reaches significance under a paired exact test, and the human-gold intervals are wide (Appendix G), so we report descriptive differences on these benchmarks; the samples do not establish population-level differences between the architecture variants. Those intervals and the cell-level paired tests also treat the 55 pilot cells as independent although they are nested in five papers, and the calibration rules were derived on the pilot they are scored on, so the calibration lift is an in-sample result whose size on unseen papers is untested. The parallel-trends criterion is likewise a reporting check: ARGUS does not test whether a pre-trend test had the power to detect a violation. The real corpus is a general-economics stress test that contains no climate-policy evaluation in the narrow sense (Appendix H); a climate-policy corpus run is the natural next test. Relatedly, flaw injection certifies the auditor against the *injected* threat on a single dimension only and inherits the taxonomy's blind spots, so we treat detection and localization as necessary rather than sufficient, pairing them with the false-alarm rate and the human-gold comparison. Evidence grounding limits coverage. ARGUS emits `unknown` only when the relevance gate finds no suitable passage; partial matches still reach the judge and can produce over-severe verdicts. Finally, ARGUS screens whether reported evidence supports a study's identification assumptions and does not adjudicate the true causal effect, which is never observed. Its outputs surface risk for expert review and do not replace it.

Acknowledgments

We thank Yan Li and Miao Zhang, the two annotators of the pilot in §5.4, who read the five papers, labelled them independently and reconciled their differences on their own time, and the anonymous reviewers, whose comments prompted the uncertainty quantification in Appendix G, the account of the retrieval mechanism in Appendix D, and the analysis in §5.3 of why the system abstains on short fixtures yet flags missing evidence on full papers. The icons in Figures 1 and 2 are from Flaticon.com.

References

Joshua D. Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Joshua D. Angrist and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Susan Athey and Guido W. Imbens. 2017. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249–275.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88.

Abel Brodeur, Nikolai Cook, and Anthony Heyes. 2020. Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–3660.

Brantly Callaway and Pedro H. C. Sant'Anna. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Clément de Chaisemartin and Xavier D'Haultfœuille. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Andrew Gelman and Eric Loken. 2014. The statistical crisis in science. *American Scientist*, 102(6):460–465.

Andrew Goodman-Bacon. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 375–385.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *International Conference on Learning Representations (ICLR)*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (TMLR)*.

Douwe Kiela, Max Bartolo, Yixin Nie, et al. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Edward E. Leamer. 1983. Let's take the con out of econometrics. *American Economic Review*, 73(1):31–43.

Donggyu Lee, Hyeok Yun, Meeyoung Cha, Sungwon Park, Sangyoon Park, and Jihee Kim. 2025. EconCausal: A context-aware economic reasoning benchmark for large language models. *Preprint*, arXiv:2510.07231.

Percy Liang, Rishi Bommasani, Tony Lee, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Bruce D. Meyer. 1995. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–161.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Ashesh Rambachan and Jonathan Roth. 2023. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Jonathan Roth. 2022. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322.

Jonathan Roth, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Ayush Sawarni, Jiyuan Tan, and Vasilis Syrgkanis. 2026. CausalReasoningBenchmark: A real-world benchmark for disentangled evaluation of causal identification and estimation. *Preprint*, arXiv:2602.20571.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based

question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931.

Raphael Silberzahn, Eric L. Uhlmann, Daniel P. Martin, et al. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Dominik Stammbach, 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.

Liyang Sun and Sarah Abraham. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models. In *AAAI 2024 Workshop "Are Large Language Models Simply Causal Parrots?"*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research (TMLR)*.

Yonghong Zhang, Ricardo Correia, Isabel M. Parra, and Yong Xie. 2026a. CausalVerify: An execution-grounded benchmark for LLM causal inference workflows. *Preprint*, arXiv:2609.07944.

Yonghong Zhang, Shadi Motaali, Vu Phong Dinh, Avin Piroutiniya, Jorge E. López de Vergara, Luis de Pedro, Ricardo Correia, Isabel M. Parra, and Yong Xie. 2026b. The double measurement confound in agent benchmarks: De-scaffolding, ground-truth scoring, and reliability beyond the mean. *Preprint*, arXiv:2609.09218.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Yu Zhou, Xingyu Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. 2026. CausalBN-Bench: A comprehensive benchmark for causal learning capability of LLMs. *IEEE Transactions on Artificial Intelligence*. Early access.

A Domain Motivation and Design Rationale

A.1 Why climate-policy DID is a demanding audit setting

Climate-policy evaluations can raise several identification challenges at once. Policy announcement, legal adoption, and actual implementation may occur at different dates, creating scope for anticipation. Carbon markets and environmental regulations may be introduced at different times across jurisdictions, so that heterogeneous and staggered treatment effects need to be handled. Emissions, investment, and production responses can spill across regional borders, challenging no-interference assumptions. Climate policies can overlap with industrial, energy, and air-pollution regulations, complicating the exclusion of concurrent interventions. Policy effects may emerge with long and heterogeneous lags. Emissions and environmental outcomes may be assembled from administrative, modelled, or remotely sensed sources whose measurement properties vary across places and time. These features do not make DID unique to climate policy; they motivate a rubric that makes the evidence for each design choice explicit.

A.2 Identification auditing and research reliability

ARGUS addresses one component of a broader research-reliability problem: published conclusions can vary with specification choices, undisclosed analytic flexibility, and researcher degrees of freedom (Leamer, 1983; Ioannidis, 2005; Simmons et al., 2011; Gelman and Loken, 2014; Silberzahn et al., 2018; Brodeur et al., 2020; Botvinik-Nezer et al., 2020). This is the identification-credibility component of the broader research-design or "credibility revolution" program in applied econometrics (Angrist and Pischke, 2010; Athey and Imbens, 2017), which increasingly assesses how sensitive conclusions are to departures from parallel trends and cautions that an insignificant pre-trend test does not validate the assumption (Rambachan and Roth, 2023; Roth, 2022). ARGUS does not attempt to audit every

Table 3: Each injected flaw, its target dimension, and its methodological basis in the literature; the reporting requirement the flaw violates is the rubric's operationalization.

flaw	dimension	methodological basis
pretrend divergence	parallel trends	(Roth, 2022; Roth et al., 2023; Rambachan and Roth, 2023)
anticipation effect	no anticipation	(Callaway and Sant'Anna, 2021; Roth et al., 2023)
forbidden comparison	staggered timing	(Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021)
spillover contamination	SUTVA / spillovers	(Athey and Imbens, 2017; Roth et al., 2023)
noncomparable controls	control group	(Callaway and Sant'Anna, 2021; Athey and Imbens, 2017)
bad control	specification	(Angrist and Pischke, 2009)
understated standard errors	inference	(Bertrand et al., 2004)
cherry-picked window	sample period	(Simmons et al., 2011)
confounding policy	concurrent policies	(Meyer, 1995)
missing placebo	placebo / robustness	(Athey and Imbens, 2017)
measurement break	data measurement	(Meyer, 1995)

component of an empirical paper; it focuses on the assumptions that support a causal interpretation.

A.3 Why the pipeline is bounded

Open-ended agents may choose different tools and reasoning paths across runs, making their outputs difficult to reproduce and compare. Large language models are increasingly used for evaluative and reviewing tasks (Zheng et al., 2023; Liang et al., 2024; Lu et al., 2024); for an audit, a fixed procedure is easier to inspect than an open-ended agent. ARGUS therefore fixes the pipeline order in code and confines model-driven decisions to evidence retrieval and adequacy assessment. Each model call operates under a fixed tool set, a hard step budget, and a logged trace. This design trades autonomy for inspectability and controlled evaluation under flaw injection; it fixes the sequence of operations, not the model's outputs.

B Flaw Taxonomy Provenance

The rubric, taxonomy, and injector are co-designed, but the catalogue of threats is not ours: each injected flaw draws on a threat or a reporting practice discussed in the DID and research-design literature (Table 3). What we authored is how each becomes a missing or flawed reported check in the rubric, and the injector's prose for it, which is why we describe synthetic detection as internal consistency under known threats and treat independently authored flaws as future work.

Table 4: Compute-graded ablation on the 33-variant benchmark (gpt-4o). det / fa / loc = detection, false alarm, localization. Calls per paper are the nominal maximum: when the gate finds nothing relevant the adequacy call is skipped, and retries are not counted. Localization takes the top-ranked flagged dimension; ties in risk level are broken by the fixed rubric order.

arm	calls/paper	det	fa	loc
full-paper single pass	1	0.727	0.000	0.727
per-dimension, no retrieval	11	0.879	0.000	0.879
two-stage (retrieval + gate)	22	0.747	0.091	0.657

C Ablation Matrix and Cross-Model Panel

Table 4 grades compute on the 33-variant benchmark: per-dimension prompting (one adequacy call per dimension) detects more variants than a single joint call, and the retrieval-and-gate layer converts omission detection into abstention. Table 5 ports the unchanged pipeline across providers: commission detection is uniformly high, while false-alarm and localization profiles vary sharply by model. Cross-provider decoding defaults differ (newer reasoning models reject an explicit temperature), so these are single runs. Table 6 shows where the operating points diverge: how each model treats a deleted section, how often it already rates the unmodified fixture medium or high, and how often its injected-variant verdict matches gpt-4o's.

On real papers the gate's behavioural contrast sharpens. This run used a second single-pass implementation, distinct from the strict-schema arm of Tables 1 and 4: its prompt permits `unknown` with the instruction to "use sparingly" (Appendix E). Over the 25 papers common to both runs (275 cells) it chose `unknown` once (0.4%; a model-written abstention with a rationale, not a harness default), where the gated pipeline abstains mechanically on 38.9%. The gated pipeline's `unknown` cells become low (1.8% → 14.9%) or medium (14.2% → 43.6%) judgements while the high-risk share stays comparable (gated 45.1%, single pass 41.1%). These shares describe behaviour (a discretionary abstention option is rarely taken, a mechanical one often is); without cell-level gold they do not show which judgements are more reliable.

Table 5: Cross-model panel on the 33-variant benchmark (single run per model; gpt-4o is the three-run mean; single runs support the qualitative porting claim only). comm/omit = commission/omission detection. Intervals in Table 8.

model	det	fa	loc	comm	omit
gpt-4o	0.75	0.09	0.66	0.89	0.45
Claude Opus 4.8	0.91	0.82	0.24	1.00	0.73
Gemini 2.5 Flash	0.79	0.18	0.36	0.91	0.55
Llama 3.1 8B (local)	0.94	0.27	0.27	0.95	0.91

Table 6: Cross-model operating points on the 33-variant benchmark (single runs; gpt-4o run 1). Omission column: verdicts on the eleven deleted-section variants, in the order `unknown` / high / medium / low. Clean column: commission pairs (of 22) whose *unmodified* fixture is already rated medium or high. Last column: exact agreement with gpt-4o's injected-variant verdict.

model	omission verdicts	clean $\geq$ med	agree
gpt-4o	6 / 3 / 2 / 0	2 / 22	–
Claude Opus 4.8	3 / 7 / 1 / 0	18 / 22	20 / 33
Gemini 2.5 Flash	4 / 3 / 3 / 1	4 / 22	26 / 33
Llama 3.1 8B (local)	1 / 10 / 0 / 0	6 / 22	19 / 33

D Retrieval Details

The only corpus queried is the paper under audit; there is no external corpus and no embedding index. Papers are ingested through a light markdown convention (`##` section headings plus `@figure`/`@table` caption lines), and the searchable blocks are the title, abstract, each section, and each figure or table caption. On the LLM path retrieval is section-level and lexical: each dimension has a fixed query pack of 11–19 distinct terms (e.g. *pre-trend, event study, lead coefficients* for parallel trends), a block scores by the summed frequency of its matched terms, and the top k=4 blocks are kept, each truncated to 2,000 characters. The relevance gate sees the first 1,200 characters of each candidate and labels it *high*, *partial*, or *none*; retrieval quality is *good* if any candidate is high, *weak* if only partial candidates remain, and *failed* otherwise. A failed dimension is scored `unknown` without an adequacy call; weak and good dimensions pass the kept blocks, in full, to the adequacy judge together with the rubric's assumption, implication, and expected evidence. The keyword baseline instead searches paragraph chunks capped at 900 characters (sentence-split when longer) with a single query string and top k=5, scoring positive and negative sentinel phrases. With `max_steps`=1 there is no iterative exploration, so breadth is the top-k cap and depth is one round. Chunk size, k, and embedding retrieval were not swept. The oracle-retrieval arm of §5.2 is a diagnostic comparison in which the target section is supplied directly and the gate is bypassed; it is not a formal upper bound on retrieval performance. Table 7 shows how the gate's three outcomes map to risk on the pilot-gold cells, the mechanism behind the fixture/real-paper asymmetry of §5.3.

Table 7: Retrieval quality versus emitted risk on the 55 pilot-gold cells (re-run recording retrieval quality and evidence status). Failed retrieval always abstains; weak retrieval almost always becomes a substantive *high*, 19 of those 20 with `evidence_status=missing`.

retrieval quality	`unknown`	high	medium	low
failed (22)	22	0	0	0
weak (22)	0	20	2	0
good (11)	0	5	5	1

E Prompts

All gpt-4o calls run at temperature 0 (other providers: Appendix F); all but the real-paper single-pass variant at the end of this appendix use a strict JSON schema. Angle-bracketed fields are substituted from the rubric entry or the retrieved passages; nothing else varies across dimensions or papers. The per-dimension no-retrieval arm of §5.2 reuses the adequacy prompt below verbatim, passing the whole paper as the evidence items instead of retrieved sections: the adequacy prompt is unchanged, but retrieval and the relevance gate are bypassed.

Relevance gate (system).

```
You are screening retrieved passages
for a difference-in-differences (DID)
identification audit.  For ONE
identification dimension, decide whether
each passage is actually evidence about
THAT dimension's assumption.  This is a
relevance check, NOT a quality or risk
judgement.  Mark a passage 'high' if it
directly concerns the assumption (e.g.
an event-study / pre-trend discussion
for parallel trends), 'partial' if it
touches it indirectly, and 'none' if it
is about something else.  Be strict:  a
passage that merely shares vocabulary
but is about a different topic is
'none'.
```

Relevance gate (user).

```
Identification dimension: <name>
Assumption: <assumption>
Testable implication: <implication>

Candidate passages:
[0] (<source>) <passage text, first 1200 characters>
[1] (<source>) <...>

Classify each passage's relevance to this dimension.
```

Schema: one `{index, relevance, reason}` object per passage, with `relevance` in `{high, partial, none}`.

Adequacy judge (system).

```
You audit the causal-IDENTIFICATION credibility of difference-in-differences (DID) studies in environmental policy evaluation. You do NOT judge whether the estimated effect is true -- the counterfactual is never observed. For ONE identification dimension, judge whether the evidence the paper reports is ADEQUATE to support the assumption.

Crucial distinction: evidence PRESENCE is not evidence ADEQUACY. The mere mention of a topic (an event study, a balance table, 'placebo') does not make the assumption supported -- assess whether the reported evidence actually establishes the testable implication, and whether it reveals a flaw.

Return a risk level and an evidence-status label:
  risk: low | medium | high
  evidence_status: sufficient | partial | missing | flawed
Map roughly: sufficient->low, partial->medium, missing->high, flawed->high. Be conservative: if the evidence is absent, thin, or describes a flawed design, do not call it supported. Cite the specific evidence (by source) your judgement rests on. Keep the rationale to one or two sentences.
```

Adequacy judge (user).

```
Identification dimension: <name>
Assumption: <assumption>
Testable implication: <implication>
Evidence a credible paper would report: <expected evidence>

Evidence retrieved from THIS paper:
- [<source>] <passage text>

Judge the adequacy of this evidence for the assumption.
```

When the gate keeps nothing the evidence block reads `(no evidence was retrieved for this dimension)`, but that path is unreachable in the deployed pipeline: a failed gate short-circuits to `unknown` without a model call (§5.3).

Schema: `{risk, evidence_status, rationale, cited_evidence}`.

Single-pass ablation, synthetic benchmarks (system and user).

```
You audit the causal-identification credibility of a difference-in-differences (DID) study. You do NOT judge whether the estimated effect is true. For EACH of the eleven identification dimensions listed, judge whether the evidence the paper reports is ADEQUATE to support that assumption. Evidence presence is not evidence adequacy. Return one judgement per dimension.
```

```
Identification dimensions to judge:
- <dimension id>: <assumption> (expect: <expected evidence>)
  [... eleven lines ...]

THE PAPER:
[<source>]
<full section text>

Judge the adequacy of the reported evidence for EVERY dimension.
```

Schema: a `judgements` array of `{dimension, risk, evidence_status, rationale}`, with `dimension` restricted to the eleven rubric ids and `risk` to `low | medium | high`; this arm cannot answer `unknown`. It produced the single-pass rows of Tables 1 and 4.

Single-pass variant, real-paper run (system, abridged to the risk block).

```
For each dimension output one risk level:
  low : adequate, credible evidence is reported.
  medium : evidence is partial, indirect, or weak.
  high : the required evidence is absent, or present but flawed.
  unknown : the text gives no basis to judge either way (use sparingly).

Also output evidence_status in {sufficient, partial, missing, flawed} (roughly sufficient->low, partial->medium, missing/flawed->high). Be conservative: absent, thin, or flawed evidence is not 'supported'.
```

The real-paper contrast of Appendix C used this second implementation (`single_pass_assessor.py`). Its preamble restates the adequacy standard of the judge prompt above, it receives the whole paper with the rubric, and it returns free JSON keyed by dimension id in which a missing or malformed cell is coerced to `unknown`.

F Reproducibility Details

Code, configuration, prompts, committed run files, the frozen pilot inputs, the annotation protocol and the locked gold are available at https://github.com/yonghongzhang-io/ARGUS.

All gpt-4o stages use temperature 0 with `max_steps`=1; newer reasoning models (Claude Opus 4.8) reject an explicit temperature and run at provider defaults. The identification rubric, flaw taxonomy, and 33-variant catalogue are declarative configuration files in the code base, and the keyword baseline plus all leak-freedom (non-circularity) checks are deterministic and run without model access. Eleven-flaw numbers are July-2026 runs of the current runner; detection replicates exactly across the `gpt-4o` alias and the dated `gpt-4o-2024-11-20` snapshot (localization varies between runs, 0.455–0.727, and the snapshot run rates one clean dimension medium, a false alarm of 0.09 where the reported run has none). A June-2026 run of the same code had scored two-stage 1.000/0.909 and single-pass 0.455; neither value was reproduced in July on the alias *or* the dated snapshot. The available records do not establish the cause of this difference, so we report the later runs with their dates and model identifiers. The 33-variant benchmark was run three times end-to-end (roughly 750 model calls per run, ∼20 minutes and ∼US$8 per gpt-4o run; the local Llama run took ∼67 minutes on a laptop at zero API cost); binary detection agreed on 32/33 variants across the three runs and target-dimension risk labels on 31/33, while the clean audit flagged a different dimension in run 1 than in runs 2 and 3, and Table 2 reports the three-run mean. The two-call-per-dimension pipeline audits one paper for roughly US$0.25 of gpt-4o usage.

G Uncertainty Quantification

Table 8 gives Wilson 95% intervals for the main proportions reported in the main text and exact tests between arms evaluated on the same items; Table 9 gives cell-bootstrap intervals (5,000 resamples) for the human-gold agreement metrics. At n=11 the keyword-versus-two-stage gap is marginal under the paired test on the same flaws (7 vs 1 discordant, p=0.07; an unpaired Fisher test, which ignores the pairing, gives p=0.03); at n=33 no pairwise difference between the single-pass, per-dimension, two-stage, and oracle arms reaches significance. For the calibration lift the comparison is paired, so overlap of the intervals in Table 9 is not the test. On the 33 answered cells the weak-retrieval rule alone turns seventeen wrong cells exact and none the other way (exact McNemar p<0.001); all four rules turn twenty (p<0.001). Those tests treat cells as independent, yet the cells are nested in five papers. By paper, the single rule improves exact agreement in all five papers, as do the four rules (sign test p=0.0625 in each case, the smallest value five papers allow). Neither analysis removes the bias of deriving the rules on the pilot they are scored on, and both count agreement with a gold that is 84% *medium*, which a rule that demotes to *medium* matches by construction. We therefore claim a consistent in-sample direction, not a demonstrated out-of-sample gain.

Tables 8 and 9 are computed from committed files by `uncertainty.py` and `calibration_recheck.py` in `experiments/ablations/`; the pilot cells they read are frozen with checksums in `experiments/annotation/pilot_frozen/` and the locked gold in `experiments/annotation/gold_final/`.

H The Real-Paper Corpus

The 26 papers audited in §5.3 are those tagged difference-in-differences in the corpus of 259 economics papers released with CausalVerify (Zhang et al., 2026a) (Table 10; titles in Table 11). The tag is the corpus's method-family label, which we did not verify paper by paper, and it is imperfect: one entry is a literature review (below) and one identifies its effect from sibling comparisons. Nineteen entries are journal articles, one of them in a medical journal, and seven are held as National Bureau of Economic Research (NBER) working-paper versions. That corpus as released contains one byte-identical duplicate (a 2002 Clean Air Act study under two identifiers); both copies re-

Table 8: Wilson 95% confidence intervals (CI) for reported rates (top) and exact tests between arms (bottom; McNemar tests are paired on the same flaws or variants; discordant = detected by the first arm only / by the second only). False-alarm rows give counts only: every variant is paired with the same clean fixture, so they are repeated readings of one clean audit of eleven dimensions, not independent trials, and an interval would overstate what they show. Detection intervals describe this fixed test set (33 variants nested in eleven dimensions of one fixture), not real papers at large.

benchmark	quantity	k/n	95% CI
11 flaws	keyword detection	2/11	0.05–0.48
	single-pass detection	9/11	0.52–0.95
	per-dimension detection	11/11	0.74–1.00
	two-stage detection	8/11	0.43–0.90
	two-stage false alarm	0/11	–
33 variants	single-pass detection	24/33	0.56–0.85
	per-dimension detection	29/33	0.73–0.95
	two-stage detection (run 1)	25/33	0.59–0.87
	commission	20/22	0.72–0.97
	omission	5/11	0.21–0.72
	two-stage false alarm	3/33	–
	oracle-retrieval detection	30/33	0.76–0.97
cross-model	Opus 4.8 detection	30/33	0.76–0.97
	Opus 4.8 false alarm	27/33	–
	Gemini 2.5 Flash detection	26/33	0.62–0.89
	Gemini 2.5 Flash false alarm	6/33	–
	Llama 3.1 8B detection	31/33	0.80–0.98
	Llama 3.1 8B false alarm	9/33	–
test	**arms**	**discordant**	p
Fisher	keyword vs two-stage (11)	–	0.030
McNemar	keyword vs two-stage (11)	1 / 7	0.070
McNemar	single-pass vs per-dim. (33)	1 / 6	0.125
McNemar	per-dim. vs two-stage (33)	5 / 1	0.219
McNemar	single-pass vs two-stage (33)	5 / 6	1.000
McNemar	two-stage vs oracle (33)	1 / 6	0.125

ceived identical audits on all eleven dimensions from both assessors, and we count the paper once, so every count reads 26 papers and 286 cells. One further entry is a review of the school-spending literature rather than a primary DID study; we retain and note it. By topic, the corpus spans finance and banking (5), health (4), education (4), environment and energy (4), international trade (3), labour (3), crime and social policy (2), and industrial location (1); the environment-and-energy subset covers the industrial-activity effects of environmental regulation, air quality and early-life mortality, nonlinear electricity pricing, and gasoline prices and vehicle choice. None of the 26 is a climate-policy evaluation in the narrow sense (a carbon price, an emissions trading scheme, or an emission standard): the real corpus is a general-economics stress test of the audit architecture, and the climate-specific content of the paper is the rubric instantiation of §3 and Appendix A; the synthetic fixture of §4 is a generic environmental-policy study (a China emissions-trading fixture ships with the code but produced none of the numbers reported here). The environment-and-energy subset has an `unknown` share of 0.43 (44 cells), against 0.39 in the remaining papers (242 cells); the corresponding high-risk shares are 0.34 and 0.46. With four papers in the subset, this comparison does not establish whether audit behaviour differs by domain.

Table 9: Cell-bootstrap 95% intervals for agreement with the reconciled gold (answered cells only; quadratic-weighted κ; 5,000 resamples). Rule 1 is the weak-retrieval rule; rules 1–4 add the three single-cell rules of Appendix J. Resampling cells ignores their nesting in five papers and may understate uncertainty.

policy	n	exact agreement	weighted κ
before	33	0.24 (0.09–0.39)	0.08 (−0.05–0.25)
rule 1, demote→med	33	0.76 (0.61–0.88)	0.23 (−0.22–0.63)
rule 1, abstain→unk	13	0.62 (0.38–0.85)	0.29 (−0.18–0.77)
rules 1–4, demote→med	33	0.85 (0.73–0.97)	0.22 (−0.42–0.79)
rules 1–4, abstain→unk	13	0.85 (0.62–1.00)	0.32 (−0.52–1.00)

Table 10: Venue of the 26 audited papers, as recorded for the version held in the source corpus. Two entries carry a corrupted venue field in the corpus index; they are counted under the journal that published them (Journal of Political Economy, 2003; Review of Economic Studies, 2002). AEJ: American Economic Journal; JF: Journal of Finance; REStat: Review of Economics and Statistics; REStud: Review of Economic Studies.

venue / version	papers
American Economic Review	4
Quarterly Journal of Economics	3
Journal of Political Economy	2
Journal of Financial Economics	2
AEJ: Applied Economics	2
JF / REStat / REStud / J. Monetary Econ. / J. Int. Econ. (one each)	5
New England Journal of Medicine	1
NBER working-paper versions	7

I Qualitative Audit Examples

A complete audit report. Table 12 shows ARGUS's actual output on the synthetic fixture with the *anticipation* flaw injected (§4): the report flags exactly the damaged dimension as high, giving the anticipation-gap rationale, keeps nine supported dimensions at low, and shows a mild over-severity trace on *inference* (the fixture does not state the number of clusters), the same bias the human-gold pilot quantifies. This is the dimension-level, evidence-cited report a reviewer receives.

Table 11: The 26 audited studies, by title as indexed in the source corpus (the duplicated entry is listed once).

The Effect of the Banking Panic on the Supply of Credit to the Corporate Sector
The Effects of State-Level Banking Competition on Innovation
Do Credit Market Shocks Affect the Real Economy? Quasi-Experimental Evidence from the Great Recession and 'Normal' Economic Times
Cyclicality of Credit Supply: Firm Level Evidence
Does Stock Liquidity Enhance or Impede Firm Innovation?
Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania
Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties
Hospital Ownership and Public Medical Spending
The Impact of Public Disclosure of Health Outcomes on Patient and Social Welfare
Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania, and Connecticut
Mortality and Access to Care among Adults after State Medicaid Expansions
Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start
The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures
Legalized Abortion and Crime
Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing
Estimates of the Impact of Crime Risk on Property Values from Megan's Laws
Effects of Liberalized Trade on Plant Productivity
The Surprisingly Swift Decline of U.S. Manufacturing Employment
Factor Immobility and Trade: Poverty in India
Trade Liberalization and Industry Wage Premia: Evidence from Colombia
The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders
Air Quality and Early-Life Mortality: Evidence from Indonesia's Wildfires
Does School Spending Matter? The New Literature on an Old Question
The Evolution of Black/White Earnings: 1960–1980
The Effects of Class Size on Student Achievement: New Evidence from Population Variation
Are Consumers Myopic? Evidence from New and Used Car Purchases

A grounded judgement on a real paper. On a trade-and-wages paper ARGUS retrieves the employment-results section and flags *inference* as high because "the paper only mentions heteroskedastic-consistent standard errors and does not provide information on clustering or the number of clusters": a specific econometric observation, not a keyword hit. Elsewhere it distinguishes a *related but different* reported check (an alternative-specification section) from the placebo evidence the rubric specifies for that dimension.

J Supplementary Figures and Tables

Pilot-gold construction (§5.4). The protocol and analysis plan were written and committed before any label was returned (`annotation/human_pilot/PROTOCOL.md`); each returned workbook is recorded by SHA-256 on receipt, and the gold is locked once (`experiments/annotation/gold_final/`). The two annotators, a student at the first author's university and a member of a vocational college in China, neither an author nor involved in building ARGUS, read the five PDFs themselves, used no AI tool, saw no ARGUS output, did not confer before returning, and signed to that effect in the workbook. The five papers are `paper_01`, `03`, `07`, `08` and `10` of the repository's corpus manifest, chosen in June 2026 before any label existed. Independent labels: 55 cells each; applicability agrees on 49, with one cell rated not applicable by one annotator; of the 54 cells both rated, 45 agree exactly (Cohen's $\kappa = 0.56$; quadratic-weighted $\kappa = 0.62$), and all nine risk disagreements are a single severity step apart. The 13 cells that differed in applicability or risk were settled by the two annotators together, all 13 as agreed decisions, 9 of them under an analogue reading of the dimension; no third person was needed. Each annotator's verbatim quotes were checked against the PDFs: 50/55 and 49/55 occur verbatim, the rest as partial matches, one cell each not found. Three deviations are logged: one annotator first returned a paper sheet pasted from another paper and corrected it herself within sixteen minutes; the file first recorded for her had been re-saved on the first author's computer (how is not known) and differed from her own last version in eleven evidence-location cells and one rationale, in no label, quote or confidence, so the record was rebuilt from her own file and none of the twelve cells enters the analysis; and the other annotator first returned the workbook without the sign-off sheet and re-sent it signed.

Table 12: ARGUS audit report (abridged) on the fixture with the injected anticipation flaw. Rationales are abridged from the evidence-cited report; the injected dimension is the one flagged high.

dimension	risk	rationale (abridged)
no anticipation	**high**	outcomes began changing between the policy announcement and its rollout, undermining no-anticipation
inference	medium	clustering and wild-cluster bootstrap reported, but the number of clusters is not stated
parallel trends	low	event-study with null pre-period coefficients and placebo tests
staggered timing	low	heterogeneity-robust estimators with a Goodman-Bacon decomposition
7 further dimensions	low	supporting evidence retrieved and judged adequate

Change from the reviewed version. The reviewed version described two annotators and an adjudication round. No record of how those la-

bels were produced survives: the first author's recollection is that two people produced them with LLM assistance and sent them to the first author, who entered them; there is no file, sign-off or adjudication record of theirs, so the reviewed description was not supported by any record, and the labels are not treated as an annotation. They are kept, checksummed, in `experiments/annotation/pilot_frozen/` and are replaced here. Against the human gold they agree on 31/55 cells (Cohen's $\kappa = 0.23$), being less severe on 23 (low where the gold is medium); under them ARGUS was more severe on 28/33 answered cells and exact on 5, against 25/33 and 8 now, and the reviewed calibration figures (exact agreement $0.12 \rightarrow 0.36$, 0.45 with four rules) are superseded by §5.6: the larger lift now reflects the medium-heavy gold, not the rule.

In the leave-one-paper-out check (§5.6), the demotion rule's trigger conditions recur in all five folds: among each fold's over-severe *high* cells, 79–86% have weak retrieval and 94–100% have missing evidence (computed on the rich re-run of §5.6). The folds test whether the error pattern recurs; no rule was refitted on four papers and scored on the fifth.

The calibration rules. The layer acts on *high* cells only and never raises a risk; the first matching rule applies. Rule 1 (20 cells): `retrieval_quality=weak` → *medium*, or → `unknown` under the abstain policy. Three further rules each fire on one cell and demote to *medium*: rule 2, the dimension is staggered timing and the rationale names Callaway, Sun–Abraham, Goodman-Bacon, or "modern estimator"; rule 3, the dimension is inference, `evidence_status=missing`, and the rationale mentions "cluster" or "robust standard errors"; rule 4, the dimension is concurrent policies and `retrieval_quality=good`. Rules 2–4 are post-hoc string and field rules defined after inspecting the errors on this same pilot, and all three cells they touch land on the gold label. They are not principled: naming an estimator, or the Goodman-Bacon decomposition (a diagnostic of two-way fixed-effects comparisons, not a heterogeneity-robust estimator), does not show that it was used appropriately; a rationale that mentions clustering may be reporting its absence; and rule 4 has no content condition at all (locating relevant evidence is not the evidence supporting the assumption). We treat only rule 1 as a result. The script (`calibrate_argus_v1_pilot.py`) reproduces the historical outputs on all 55 cells.

Table 13: Two-stage ARGUS vs reconciled human gold (55 cells, 5 papers) under two `unknown` policies: coverage (drop abstentions) and strict (`unknown` counts as mismatch). Pilot-scale: only 2 gold cells are *high*; the robust signal is the one-directional over-severity and the abstention rate.

`unknown` policy	n	exact	κ	wt. κ
exclude (coverage)	33	0.242	0.031	0.128
mismatch (strict)	55	0.145	0.007	–

Table 14: Deterministic calibration vs the human gold (55 cells), in-sample. Two policies for a weak-retrieval *high*: demote to *medium* (med, keep coverage) or abstain to `unknown` (unk). Rule 1 alone is the result we claim; rules 1–4 add three single-cell exploratory rules (Appendix J). Under abstention rule 1 changes no retained label: its higher agreement comes from dropping 20 answers. Pilot scale: only 2 gold cells are *high*; a constant *medium* label scores 26/33 on the answered cells, so exact agreement carries little information here. The before column is the rich re-run of §5.6, which differs from the corpus run of Table 13 in one cell (a gold-low cell moves from *medium* to *high*, which is why weighted κ is 0.08 here and 0.13 there); intervals in Table 9.

		rule 1		rules 1–4	
metric	before	med	unk	med	unk
over-severe (answ.)	25	8	5	5	2
high precision	1/25	1/5	1/5	1/2	1/2
high recall	1/2	1/2	1/2	1/2	1/2
exact agr. (answ.)	0.24	0.76	0.62	0.85	0.85
wt. κ (answ.)	0.08	0.23	0.29	0.22	0.32
answered / `unk`	33/22	33/22	13/42	33/22	13/42

The remaining material is Figure 5 (per-dimension risk on the real corpus) and Tables 13–14 with Figure 6 (agreement and calibration against the gold).

K Shared-Evidence Control

Table 1 compares pipelines that differ in the judgement policy *and* in the evidence they retrieve. This control holds the evidence fixed. Its design, thresholds, and decision rule were committed before this control was run (`SHARED_EVIDENCE_PROTOCOL.md`, commit `288e01f`, 20 September 2026, tagged in the repository; the runner, the two zero-call cells, and the model calls followed). Two evidence bundles are frozen to disk for every

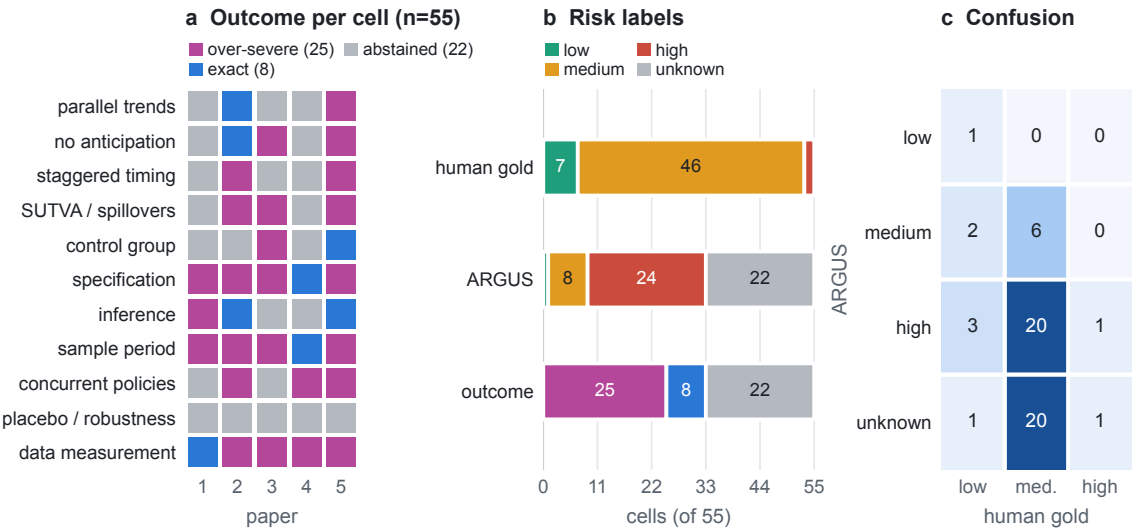


Figure 4: Two-stage ARGUS versus the reconciled gold on the 55 pilot cells (frozen protocol). **(a)** Every cell, eleven dimensions by five papers, coloured by outcome: ARGUS abstains on 22 and, of the 33 it answers, is more severe than the gold on 25, matches it on 8, and is less severe on none. **(b)** Distribution of risk labels in the gold and in ARGUS's output, and the outcome of the 55 cells in the colours of panel (a). **(c)** ARGUS-by-gold confusion matrix (counts).

Table 15: Shared-evidence control: evidence bundle (K, keyword chunks; S, ungated sections) × judgement policy (kw, keyword scorer; llm, adequacy judge). det / fa / loc as in Table 4; off = share of the ten non-target dimensions rated ≥ medium on the injected variants. The reference row is the gated pipeline on the gpt-4o alias (detection and false alarm from run 1, Table 8; localization is the three-run mean of Table 2).

	11 flaws			33 variants			
cell	det	fa	loc	det	fa	loc	off
K-kw	2/11	0/11	2/11	8/33	0/33	8/33	0.00
K-llm	10/11	1/11	8/11	31/33	3/33	22/33	0.15
S-kw	1/11	0/11	1/11	2/33	0/33	2/33	0.00
S-llm	11/11	0/11	11/11	31/33	0/33	22/33	0.16
gated two-stage	8/11	0/11	8/11	25/33	3/33	0.66	–

paper and dimension: K, the keyword pipeline's stage-2 chunks, and S, the LLM path's four section candidates with the relevance gate switched off. Each bundle is read by both policies, the keyword scorer and the adequacy judge (gpt-4o-2024-11-20, temperature 0, prompts of Appendix E verbatim). No cell is gated, so none can abstain. All 45 papers (one clean fixture, 11 flaws, 33 variants) are judged on all eleven dimensions: 990 adequacy calls, none failed. Cell K-kw reproduces the published keyword baseline exactly.

Result. With the evidence identical, the policy contrast is large under both bundles (Table 15): on K, detection rises from 2 to 10 of 11 flaws (eight gained, none lost; exact McNemar p=0.008) and from 8 to 31 of 33 variants (p<0.001); on S, from 1 to 11 and from 2 to 31. The evidence contrast at a fixed judge is small (10 vs 11 of 11; 31 vs 31 of 33). The pre-specified rule asked for a gain of at least 0.30 on the 11 flaws, the same direction on the variants, and false-alarm and clean off-target rates within 0.10 of the keyword scorer's. All three hold, the last narrowly (0.09), so the policy explanation is supported: on these fixtures the judgement policy, not the retrieval path, carries the gap in Table 1.

What it costs, and what it does not show. The gain is a trade. The judge rates 11–16% of non-target dimensions medium or higher on injected papers, where the keyword scorer flags none. On the keyword chunks the judge also rates one clean dimension (SUTVA) medium. Every false alarm in the K-llm row is that single verdict, counted once per flaw that targets the dimension: both benchmarks share one clean fixture, so the false-alarm rates here and in Tables 1 and 2 rest on one clean audit of eleven dimensions, not on 11 or 33 independent trials. The ungated S-llm cell detects more than the gated pipeline (omission variants 11/11 vs 5/11), the trade §5.2 describes: shown four sections of a short fixture, the judge notices the deleted one, where the gate abstains. The reference row was run on the gpt-4o alias, so that comparison is indicative; fixtures, sentinel lists, and injector are co-designed, and n is small.

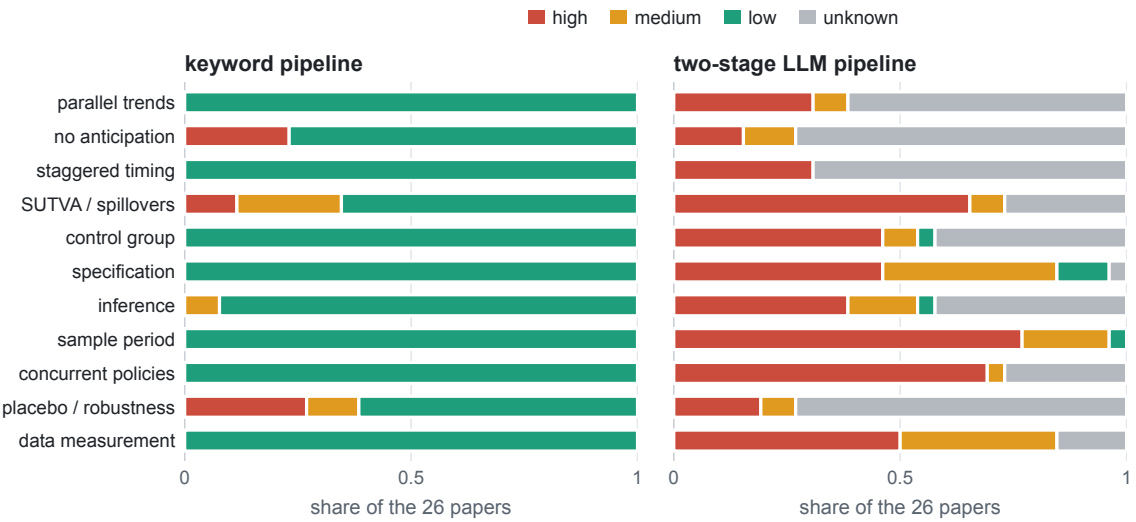


Figure 5: Per-dimension risk distribution on the 26 papers tagged DID in the source corpus. The keyword baseline (left) is almost entirely *low*; the two-stage LLM (right) is differentiated, with a large `unknown` share that marks retrieval failure rather than over-flagging.

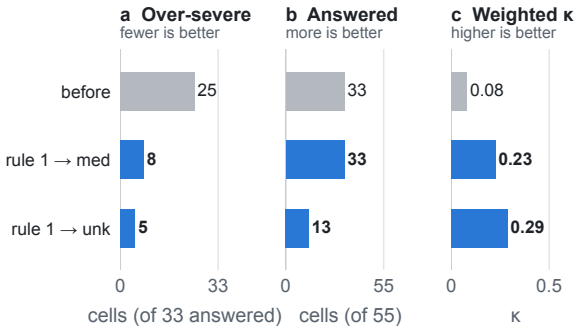


Figure 6: **Calibration trade-off: severity reduction versus coverage** against the reconciled human gold, for the weak-retrieval rule alone (in-sample; med and unk are the two policies of Table 14, which also gives the exploratory four-rule set). Demoting weak-retrieval *high*s to *medium* preserves coverage (33 answered) while cutting over-severe judgements from 25 to 8; abstaining leaves fewer over-severe cells (5) and a higher weighted κ (0.29) only by discarding 20 answers, and leaves the remaining labels unchanged. Most of the gain in exact agreement reflects a gold that is 84% *medium* (§5.6); the remaining bottleneck is evidence grounding, not severity thresholding alone.

The Energy Footprint of LLM-Based Environmental Analysis: LLMs and Domain Products

Alicia Bao[1,2,*] **Angel Hsu**[1,2,*] **Diego Manya**[1,2,*] **Ji (James) Zhang**[3,*] **Jiamian He**[3,*]
[1]The University of North Carolina at Chapel Hill
[2]Data-Driven EnviroLab
[3]Arboretica
alicia.bao@unc.edu, angel.hsu@unc.edu, diego.manya@unc.edu,
jiamian@arboretica.com, james@arboretica.com

Abstract

As large language models (LLMs) are increasingly used, understanding their energy footprint has become an important concern. The growing adoption of retrieval-augmented (RAG) based pipelines for climate-domain specific analysis raises a key question: how does the energy consumption of domain-specific RAG-based pipelines compare with that of direct generic LLM usage? Prior research has focused on standalone model calls or coarse token-based estimates, while leaving the energy implications of deployed application workflows insufficiently understood. In this paper, we assess the inference-time energy consumption of two LLM-based climate analysis chatbots (ChatNetZero and ChatNDC) compared to the generic GPT-4o-mini model. We estimate energy use under actual user queries by decomposing each workflow into retrieval, generation, and hallucination-checking components. We also test across different times of day and geographic access locations. Our results show that the energy consumption of domain-specific RAG-based pipelines depends strongly on their design. More agentic pipelines substantially increase inference-time energy use, particularly when used for additional accuracy or verification checks, although they may not yield proportional gains in response quality. This study provides a new understanding on how the design of domain-specific LLM products affects both the energy footprint and quality of output.

1 Introduction

While the rapid proliferation of LLM-powered chatbots represents a transformative shift in how information is generated and accessed, major questions remain regarding the energy-intensity and environmental footprint of utilizing these tools, particularly due to the increasing number of data centers for generative AI models training and inference (Meredith, 2023; Copley, 2025). Although data centers are estimated to only account for 1% of total global electricity demand (IEA, 2025), this percentage is projected to grow to as much as 12% in 2030 (Chen, 2025). Prior research has demonstrated that while model training is technically much more energy-intensive than inference (Kaack et al., 2022), the inference phase in aggregate is responsible for the vast majority of AI's energy consumption(O'Donnell and Crownhart, 2025). Hyperscaler AI companies like Microsoft and Google have released estimates of the water, energy and environmental impacts of AI usage (Elsworth et al., 2025), although critics have pointed out the lack of transparency and potential inaccuracies in these estimates (Li et al., 2025).

Here, we present an empirical end-to-end carbon emission assessment of real-world, domain-specific LLM-RAG-based products used for environmental analysis. We introduce a query-level methodology to decompose and quantify the energy consumption of RAG-based pipeline sub-components, rigorously benchmarking the relative costs of the Retrieval and the Generation stages. Ultimately, our findings yield the first quantitative framework for evaluating this complex environmental tradeoff (energy consumption vs quality), leading to actionable, data-driven architectural guidelines for designing demonstrably more sustainable generative AI systems.

2 Background

2.1 ChatNetZero and ChatNDC - Climate-domain specific chatbots

ChatNetZero.ai (CNZ) is a climate-specific chatbot, pairing a RAG-based pipeline with validation and referencing modules (Hsu et al., 2024). Heterogeneous data sources (e.g., PDFs and spreadsheets) are embedded for semantic retrieval. CNZ applies query processing that distinguishes actor-specific questions (e.g., named governments or companies) from generic questions, prioritizing

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 220–230
October 28, 2026 ©2026 Association for Computational Linguistics

detected Net Zero Tracker (Tracker, 2025) entities. Then an anti-hallucination module embeds the draft output sentence-by-sentence and verifies each claim against retrieved source chunks, excluding sentences that cannot be traced back to the underlying corpus. Finally, CNZ extends beyond narrative Q&A by transforming structured Net Zero Tracker variables into natural-language statements, enabling quantitative and counting-style queries to be handled through the same retrieval-and-grounding workflow.

ChatNDC (ChatNDC.org, beta) is a climate policy chatbot for credible analysis of countries' Nationally Determined Contributions (NDCs) under the Paris Agreement based on a curated corpus of verified climate documents reported to the United Nations Framework Convention on Climate Change (UNFCCC), country-specific scientific and policy assessments, and global syntheses on climate science, finance, mitigation, and adaptation. In addition to its general responses it also allows users to directly interrogate documents through a data catalog feature From a design perspective both ChatNetZero and ChatNDC follow a similar structure: A vector database built from curated sources queried through a RAG-based retrieval pipeline with an anti-hallucination loop to provide accurate and validated information to the user with traceable sources and citations. The main difference between the two products is the use of LLMs as part of the retrieval and anti-hallucination stages in ChatNDC using AI agents, as opposed to the direct semantic search used in ChatNetZero.

2.2 LLM energy consumption components: training, inference

For the purposes of evaluating LLM-based chatbots, operational energy consumption consists primarily of two main phases: training and inference (Luccioni et al., 2024). Training includes pre-training and any subsequent fine-tuning and is typically concentrated in short but highly compute-intensive runs (Chen, 2025). For instance, the training of GPT-4 consumed around 50 GWh of electricity (Ariyanti Sri and Suryanegara Muhammad). Similarly, while Google's Gemini series employs more energy-efficient architectures and specialized Tensor Processing Units (TPUs), the training cost for LLMs remains at the thousands of average household volume (Cottier et al., 2025).

Unlike the total training energy consumption, which is fixed once the model is deployed, the inference is continuous and scales directly with usage volume. Evidence indicates that for a model of GPT-3's scale, once daily active users reach the tens of millions, it takes a mere 43 days of continuous operation for cumulative inference energy to equal the initial training energy (Guardia et al., 2024). For popular models, this 'energy parity point' is reached even more rapidly(Luccioni et al., 2024).

Early work on the environmental cost of LLMs and AI emphasized the training phase, while more recent work has argued that inference deserves equal or greater attention (Lacoste et al., 2019; Kaack et al., 2022; Luccioni et al., 2024).

2.3 Previous research on measuring LLM inference methods

To quantify carbon emissions during the LLM inference phase, current research primarily adopts three methods:

Hardware-based Monitoring (Local Deployment): The most prevalent approach involves direct energy measurement via hardware-level sensors (GPU, RAM, CPU) or Cloud Service Provider (CSP) embedded calculators. This method is favored for its high precision and controllability within laboratory environments. Consequently, such studies (Everman et al., 2023; Dauner and Socher, 2025; Nguyen et al., 2024) typically deploy open-source models for measurement. Similarly, although prior work (Wu et al., 2025) has decomposed RAG for more precise energy measurement, it still relies on locally deployed small models, which differ from the models and deployment settings used in real-world industrial applications.

Architecture-driven Parameterized Estimation: Analytical tools, such as LLMCarbon (Faiz et al., 2024) and LLMCO_2 (Fu et al., 2025), utilize parameterized models to estimate energy consumption and carbon emissions. These techniques rely heavily on specific model architecture details to compute Floating Point Operations (FLOPs). Furthermore, to bridge the gap between theoretical FLOPs and actual consumption, the R-ICE (Sikand et al., 2025) research trains regression models within the framework, creating benchmarks that more accurately map inference carbon and energy metrics across diverse hardware configurations.

End-to-End API-based Estimation: For commercial "black-box" models, where architectural details and local deployment are limited, end-to-end API estimation has emerged as an efficient eval-

uation path. Related studies include the MELODI framework (Husom et al., 2026), and the subsequent work by Jegham et al. (Jegham et al., 2025) both of which systematically evaluate the environmental impacts of LLM inference. Jegham et al. (Jegham et al., 2025) propose a framework that estimates energy usage based on three core components: inference latency, estimated model/GPU scale, and PUE. This method is generalizable and a viable approach for assessing black-box or commercial models.

2.4 Limitations of previous research

Previous approaches to estimating LLM energy consumption show limited generalization particularly for commercial LLMs due to the lack of information on hardware which makes analytical methods rely on proprietary architectural assumptions. Although recent API-based frameworks estimate energy using latency and hardware specifications, they rely on average metrics that fail to capture the reality of Retrieval-Augmented Generation (RAG) and agentic workflows. (Jegham et al., 2025)

This study addresses these gaps by shifting the unit of analysis from the standalone model call to the deployed application workflow. We decompose RAG-based pipelines, making it possible to identify where energy is actually consumed within chatbot systems, including components that would be obscured in model-level averages such as query classification, retrieval, response generation, and hallucination checking. Also, by evaluating both energy use and response quality, our framework makes visible the trade-offs between computational cost and performance that are central to the real-world design of LLM-based applications.

3 Methodology

3.1 Dataset

To evaluate energy consumption and response quality under realistic usage conditions, we constructed a dataset of 102 climate related test questions reflecting the kinds of queries users are likely to ask in deployed climate-LLM applications. The dataset was manually curated to cover both direct factual retrieval and higher-order analytical tasks. We classified the questions following Bloom Taxonomy (Ormell, 1974): Knowledge, Comprehension, Application, Analysis, Evaluation and Creation (with 48 questions of the Knowledge class and approximately 10 of each of the other classes). See B for example questions. We also performed manual evaluation of factual accuracy and embellishment (i.e., an enhanced but factually ungrounded LLM-generated statement) for 20 questions for each product (ChatNetZero and ChatNDC) and GPT 4o-Mini as a reference model.

3.2 Energy Consumption Estimation

3.2.1 ChatNetZero & ChatNDC structure

Unlike a direct call for LLM API, domain RAG-based pipelines have different structures. CNZ uses a simple RAG-based pipeline as shown in Figure 3.1. It mainly contains one retrieval, one LLM call with a fine-tuned GPT 4o-mini model, and uses the pairwise cosine similarity method for hallucination check. If the cosine score is smaller than a threshold, the sentence in response will be deleted from the final output.

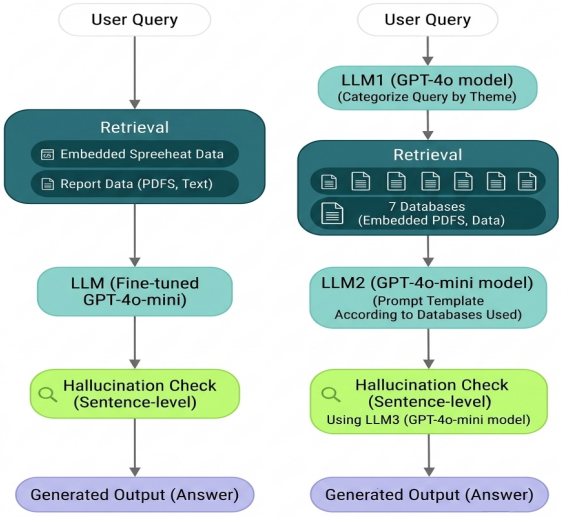


Figure 3.1: Architectures of a) ChatNetZero and b) ChatNDC.

ChatNDC is a more complex RAG-based pipeline, with two more LLM inference calls than CNZ. As shown in Figure 3.1, the system first uses an LLM inference (GPT-4o) to categorize the user query by theme, which determines the relevant databases for retrieval. The retrieved content is then passed to the second LLM inference call (GPT-4o-Mini) with a database-aware prompt template to generate the response. Finally, a third LLM inference call (GPT-4o-Mini) performs a sentence-level hallucination check to validate the generated content before outputting the final answer. Compared to the CNZ cosine similarity method, which

relies on a static threshold value, ChatNDC leverages an LLM to automatically verify the reliability of generated content against the retrieved chunks.

Thus, for the RAG-based pipelines, the total energy consumption (E_{total}) is modeled by aggregating the energy costs across three main inference parts of the RAG-based pipelines:

Retrieval: retrieve the documents in the vector database by query.

LLM Inference: API request to completion of the generated answer.

Hallucination Check: pairwise algorithm (CNZ) or LLM for the sentence-level consistency verification (ChatNDC).

The total energy consumption is represented by:

$$E_{total} = E_{retrieval} + E_{inference} + E_{hc} \quad (1)$$

More details on the variance in unconstrained models versus constrained models like CNZ, ChatNDC and GPT-4o-mini (200) (Appendix A).

3.2.1 Experiment design

To quantify the inference-time and energy consumption of different LLM-based chatbot pipelines, we decompose the computational steps and measure execution time within each workflow. We also evaluate their performance under variation in time of day. This design allows us to generate a more representative estimate of average energy consumption by accounting for fluctuations in latency associated with network conditions and server load. Detailed energy-estimation procedures are described in Sections 3.2.2-3.2.4.

All experiments were conducted in the Google Colab environment using Python 3 and single-core CPU runtime. LLMs were accessed through API calls with the "temperature" parameter fixed at 0.

Because both CNZ and ChatNDC use GPT-4o-Mini as their primary generation model, we restrict comparison to four pipelines: CNZ, ChatNDC, GPT-4o-Mini, and GPT-4o-Mini (200). The GPT 4o-Mini (200) condition imposes a maximum output length of 200 words in order to provide a controlled baseline for comparison with CNZ, whose responses are inherently constrained to this approximate length. Although absolute energy consumption can vary across other model architectures and hosting environments, holding the underlying model constant improves comparability across our experimental conditions.

The experiment consists of two complementary parts, both based on the same evaluation dataset. The first part focuses on energy estimation robustness. To account for variation in network latency and API server load, each pipeline was executed repeatedly across four runs: one run at a randomly selected time and three runs during fixed time windows in the morning (08:00–10:30), afternoon (14:00–16:30), and evening (20:00–22:30). All runs originated from the Netherlands. For each run, we recorded complete execution time and token usage and converted these measurements into estimated energy consumption. Query and response tokens were counted using the tiktoken library (noa, 2026) configured for GPT-4o-Mini.

Finally, we combine estimated cumulative energy consumption with response quality evaluation to examine the trade-offs between computational cost and performance across systems. By jointly examining energy use, token consumption and answer quality, we assess the relative efficiency of each pipeline and identify whether additional workflow and RAG design complexity yield meaningful improvements in output quality.

3.2.2 LLM consumption estimation method

Building upon Jegham et al.'s methodology (Jegham et al., 2025), we refine their approach to measure LLM energy consumption. As shown in Equation 2, $E_{inference}$(kWh) represents the LLM energy consumption. The total inference time T, corresponds to the elapsed time from calling the API until the final full answer is obtained. W represents the total active power consumption of the hardware system. PUE (Power Usage Effectiveness) represents the data center's infrastructure overhead factor. The energy consumption is defined as follows:

$$\begin{aligned} E_{\text{inference}} &= T \cdot W \cdot \text{PUE}, \\ \text{where } T &= \left(\frac{\text{Output Length}}{\text{TPS}} + \text{Latency} \right) / 3600, \\ W &= P_{\text{GPU}} \cdot U_{\text{GPU}} + P_{\text{non-GPU}} \cdot U_{\text{non-GPU}} \end{aligned} \quad (2)$$

where

$$U_{\text{GPU}} = \frac{G \cdot D_{\text{GPU}}}{N \cdot B}, \qquad U_{\text{non-GPU}} = \frac{G \cdot D_{\text{non-GPU}}}{N \cdot B} \quad (3)$$

Here, G denotes the number of GPUs assigned per model, N the number of GPUs per node, and B is the batch size. D_{GPU} represents the assigned power draw of GPUs, expressed as a fraction of their maximum rated power, while $D_{\text{non-GPU}}$ corresponds to a conservatively estimated fixed utilization fraction for non-GPU components (e.g., CPU,

memory, storage, cooling). The PUE is the 1.12 as mentioned in the Microsoft datacenter report (Microsoft) where OpenAI holds their LLM models. The LLM parameters, as reported in the reference paper, are summarized in Table 3.1.

Model	Size	GPU	G	D_{GPU}	$D_{\text{non-GPU}}$	N	B
GPT-4o	Large	H100	8	0.6	0.5	8	8
GPT-4o-Mini	Medium	A100	4	1.2	0.5	8	8

Table 3.1: Parameters and power utilization values for the evaluated models.

Therefore we estimate W according to the Equations 2 and 3, and T is obtained from the previously outlined and we estimate the parameters of the formula, presented in the Table 3.2.

Model	U_{GPU}	$U_{\text{non-GPU}}$	P_{GPU}	$P_{\text{non-GPU}}$	W
GPT-4o	0.075	0.0625	5.6	4.6	0.7075
GPT-4o-Mini	0.075	0.03125	3.2	3.3	0.343125

Table 3.2: Energy metrics and total power calculation for the evaluated models.

3.2.3 Retrieval consumption estimation method

In the retrieval part, since all processes are executed within a Google Colab environment using a single CPU core (typically an Intel(R)-Xeon(R)-CPU@2.20GHz), the retrieval energy consumption is approximated as the energy draw of a single CPU core. The official specifications for this Intel(R)-Xeon(R)-CPU series (Intel) list several variants with TDP values ranging from 150W to 330W, with a number of cores ranging from 12 to 44, but the Colab platform does not specify the exact sub-model utilized during the session. For cross-verification, we adopt a power consumption standard of about 8.5W per core based on LLM-Carbon data to serve as our primary calculation benchmark. Furthermore, to accurately reflect the total energy footprint, including infrastructure overhead, we incorporate a PUE value of 1.09, as disclosed in the Google Data Center Efficiency Report (Google). The retrieval final energy consumption is thus determined as follows:

$$\begin{aligned} E_{\text{retrieval}}(\text{kWh}) &= T_{\text{retrieval}} \cdot W_{\text{CPU}} \cdot \text{PUE} \\ &= T_{\text{retrieval}}\,(\text{h}) \cdot 0.0085\,(\text{kW}) \cdot 1.09 \end{aligned} \tag{4}$$

3.2.4 Hallucination Check consumption estimation method

As described above in section 3.2.1, CNZ and NDC employ different hallucination-checking approaches, and these differences are reflected in how their energy consumption is estimated. In CNZ, hallucination checking is implemented through sentence-level cosine similarity between each generated sentence and the retrieved source documents. Because this process does not require an additional LLM call, its energy use is attributed to the retrieval/post-processing stage and accounted for within $E_{\text{retrieval}}$. In contrast, ChatNDC performs hallucination checking through an additional GPT-4o-Mini call; therefore, the energy consumption is estimated using the same approach as for $E_{\text{GPT-4o-mini}}$.

$$E_{\text{CNZ},\,hc}\,(\text{kWh}) = T_{hc}\,(\text{h}) \cdot 0.0085\,(\text{kW}) \cdot 1.09 \tag{5}$$

$$E_{\text{NDC},\,hc}\,(\text{kWh}) = T_{hc}\,(\text{h}) \cdot 0.343125\,(\text{kW}) \cdot 1.12 \tag{6}$$

3.3 Evaluation

To assess the answers, a Human-Expert Evaluation method (Belz and Reiter, 2006) is conducted at the statement level, treating each statement as the fundamental unit of assessment.This evaluation was focused on the questions and answers within the thematic scope of ChatNetZero and ChatNDC since we wanted to evaluate real use cases of these specialized tools, and their architectures, against their GPT equivalent. Each statement within a generated response is evaluated in two aspects: Factual Accuracy and Embellishment Prevalence.

Factual Evaluation: For statements containing verifiable claims or ground-truth-dependent information, a binary scoring system is applied. A statement receives a score of 1 (Correct) if and only if it is entirely consistent with the ground truth; otherwise, it receives a score of 0. The aggregated Factual Index for each answer is then calculated as the ratio between the number of correct factual statements and the total number of factual statements.

Embellishment Evaluation: For each answer we identified the number of Non-Factual statements, defined as statements that are not associated to a validated source. The embellishments index for each answer is calculated as the ratio between the number of non-factual statements to the total number of statements in the answer.

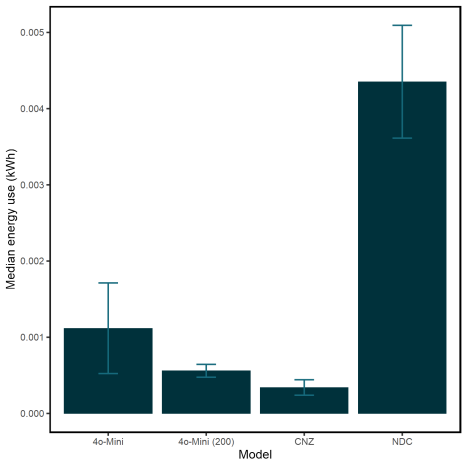


Figure 4.1: Median energy use of models per query.

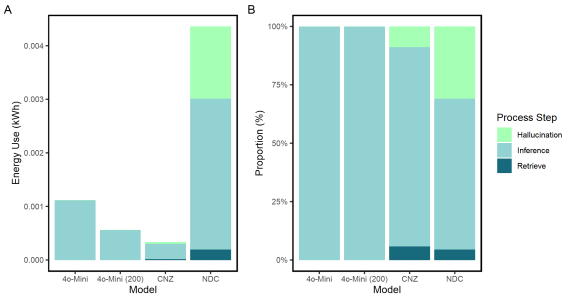


Figure 4.2: Energy use across query steps: (a) Mean energy use of models by each step of the query; (b) Proportion of energy use by step of the query.

4 Results

4.1 Energy consumption

Our results indicate that models vary in terms of their median energy consumption, with the climate-domain specific chatbots CNZ and ChatNDC not necessarily consuming less energy compared to GPT-4o-Mini. Figure 4.1 shows the median energy consumption per successful query across the four model configurations evaluated. Among all systems, ChatNDC had the highest energy use, consuming 4.53e-3 kWh per query. This consumption was substantially higher than GPT 4o-Mini, which consumed 1.13e-3 kWh, and less energy when constrained to a response of 200 words or less (4o-200: 5.0e-4 kWh). CNZ had the lowest mean energy consumption per query (4.08e-4 kWh) and the smallest variation, with a mean absolute deviation of 1.01e-4. Assuming a 100-W television, these amounts correspond to approximately 2.7 minutes, 41 seconds, 18 seconds, and 15 seconds of television operation for ChatNDC, GPT 4o-Mini, 4o-200, and CNZ, respectively. In contrast, ChatNDC showed both the highest average consumption and the largest variability, consistent with a workflow that includes multiple model calls and a more elaborate post-retrieval verification process compared to CNZ.

Figure 4.2 decomposes total query energy into retrieval, inference, and hallucination-checking components. Across all models, inference accounts for the largest share of total energy use. For the pure GPT baselines, energy consumption was almost entirely attributable to inference, since we were unable to assess estimates of energy consumption for the retrieval stages. In the case of CNZ, inference also remained a dominant component, while retrieval and its cosine-similarity-based hallucination check contributed only a small share of the total energy use. This pattern is consistent with CNZ's lightweight verification design, which does not require an additional LLM call like ChatNDC.

ChatNDC exhibited a markedly different energy profile from CNZ. Although inference remained the largest single component of total energy use in both systems, hallucination checking accounted for a substantially larger share in ChatNDC. Specifically, ChatNDC's hallucination-checking stage represented 30.9% of total energy use, compared with only 8.4% in CNZ. ChatNDC's agentic-based hallucination check consumed approximately 45 times more energy than CNZ's lighter-weight algorithm. Retrieval, by contrast, remained the smallest energy component in both RAG-based pipelines, and was similar in proportional terms across CNZ and ChatNDC. These results show that the main difference in energy use between the two RAG-based pipelines lies not in retrieval, but in the much more energy-intensive validation stage introduced by LLM-based hallucination checking.

The retrieval and hallucination steps observed in CNZ and NDC queries are part of the additional processes in their architectures. Although GPT models can search the internet when a topic is outside their training data, the API does not expose explicit retrieval or hallucination steps.

Exploring the relation between output token size and Energy consumption (Figure 4.3), we observe a strong relation between both variables for the answers generated by CNZ, GPT 4o-Mini and 4o-Mini (200). To the contrary the results of ChatNDC shows a weaker association between output token size and energy consumption. These results highlight not only the direct relation between these two

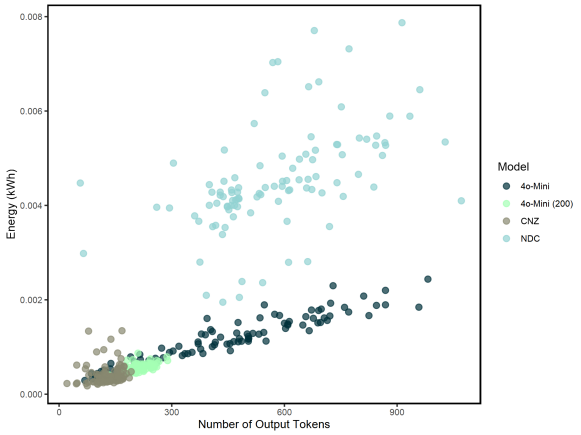


Figure 4.3: Output token vs Energy Usage by model.

variables, but also the potential effect of truncating the answers for increased energy efficiency. Finally, it also highlights that depending on the system architecture that relation can be modified such as the case of ChatNDC where output length is not strongly correlated to energy, due to the additional steps required by the model and its configuration.

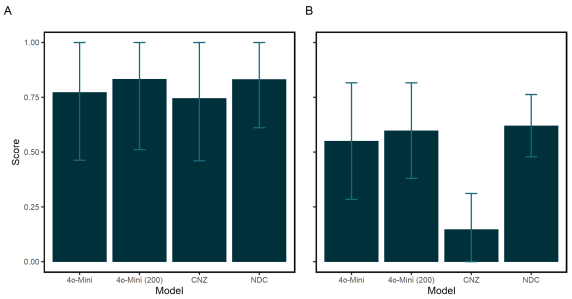


Figure 4.4: a) Mean factual score by model. b) Mean embellishment score by model.

Figure 4.4 shows that mean factual scores are broadly similar across models, with only modest differences relative to the variation across questions. By contrast, embellishment differs more clearly by system. CNZ has the lowest mean embellishment score of CNZ (M = 0.15; SD = 0.16), which is notably significantly lower than the three other models (GPT-4o-Mini (M = 0.55, SD = 0.27); GPT-4o-Mini (200) (M = 0.59, SD = 0.22), and ChatNDC (M = 0.62, SD = 0.14)). This result suggests that CNZ produces more concise and tightly grounded responses, while the other systems tend to generate more elaborate outputs without clear gains in factual performance.

Figure 4.5 reinforces this interpretation. There is no clear evidence that higher energy use is associated with higher factual scores. Instead, depending

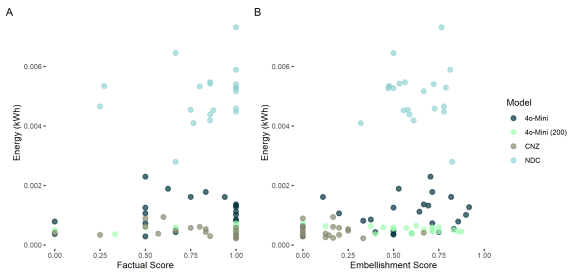


Figure 4.5: a) Energy use vs. factual score of questions, by model. b) Energy use vs. embellishment score of questions, by model.

on the output length of LLM-generated output, systems with longer outputs use substantially greater energy consumption (Adamska et al., 2026; Podder et al., 2026). In the case of ChatNDC, we found that longer outputs and subsequently higher energy consumption does not necessarily translate to higher factual performance and lower embellishment.

Figure 4.5 explores the association between embellishment, factual accuracy and output length, as measured by token size. Shorter outputs do not necessarily guarantee reduced embellishment, as observed in GPT-4o-Mini (200)'s relatively high embellishment score (score = 0.59). CNZ, however, which is prompted to keep responses under 150 words (Hsu et al., 2024), had a much lower embellishment score (0.15). Longer responses, such as those generated by GPT-4o-Mini and ChatNDC, were more often associated with higher embellishment, particularly ChatNDC. In ChatNDC's case, this result appears to stem partly from its system design, which requires responses between 400-600 words in a research-paper style. For instance, questions such as "How many NDCs has Brazil submitted" that could be answered directly in a single sentence were often followed by additional and sometimes embellished content to satisfy the target length. This result suggests that although RAG-based architectures can help reduce non-factual statements, those gains may be offset when models are required to produce unnecessarily long answers.

Complete dataset of questions and energy estimates is available in (Manya-Gutierrez et al., 2026)

5 Discussion

Our findings show the trade-off that while a RAG pipeline can improve factual grounding, it introduces additional computational steps that may increase inference-time energy use. We find that

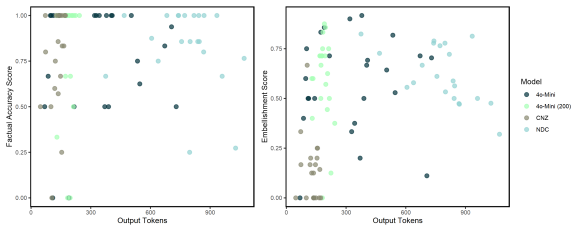


Figure 4.6: a) Output token size vs. factual score of questions, by model. b) Output token size vs. embellishment score of questions, by model.

inference-time energy use is shaped not only by the underlying model, but also by the amount of text a system is designed to generate. Prior work shows that output length is more often correlated with higher energy consumption, since each generated token requires additional inference time. Controlled experiments by Poddar et al. (Poddar et al.) show near-linear energy growth with output length and also find that, when input and output lengths are held constant, task complexity itself has only a small effect on energy use. Luccioni et al. (Luccioni et al., 2024) likewise emphasize that generative tasks are more energy-intensive than discriminative ones. In this context, our results suggest that output-length decisions are not only an energy issue but also a quality issue: systems that encourage longer or more elaborated answers may create more embellishment while increasing energy consumption.

Our comparison between CNZ and ChatNDC further shows that system design choices can have large consequences for energy consumption even when systems rely on the same base generation model. Both products are RAG-based pipelines, yet their energy profiles diverge sharply because of differences in workflow structure. In our experiments, the main distinction was not retrieval itself, but the additional inference steps introduced by ChatNDC's more agentic verification process. This result is consistent with the broader agentic RAG literature, which describes agentic pipelines as adding planning, reflection, iterative refinement, and tool use on top of standard retrieval and generation (Singh et al., 2025). Such steps may improve flexibility and robustness, but they also add inference-time compute through additional LLM calls. Our results emphasize that the energy cost of agentic RAG-based pipeline is driven less by retrieval than by post-retrieval reasoning and validation steps.

These findings point toward a more selective design strategy for future RAG-based pipelines. Rather than applying the same high-cost workflow to every query, developers may achieve better energy efficiency by routing questions according to their complexity and information requirements(Ong et al., 2025; Chen et al., 2023). For domain-specific RAG-based pipelines, our findings suggest a practical next step: simple factual retrieval questions could be handled by a lightweight pipeline, whereas open-ended, comparative, or reasoning-heavy questions could trigger a larger model or a more agentic verification stage. When developers design domain-specific RAG-based pipelines, energy should become an explicit consideration alongside factual accuracy, latency, and usability.

6 Conclusion

This study examined the inference-time energy consumption of LLM and RAG-based climate chatbots by moving beyond model-level estimates to workflow-level evaluation under real user queries. Our results show that energy use is shaped not only by the underlying model, but also by system design choices, including retrieval structure, output length, and hallucination-checking method. In particular, more agentic RAG-based pipelines can increase energy consumption without necessarily producing proportional gains in response quality in the evaluation setting examined here. It also suggests that domain-specific RAG-based pipelines are not inherently more or less energy-intensive than generic LLM use, but their energy profile depends on how the workflow is constructed. By jointly evaluating energy consumption and answer quality, this study contributes a more application-relevant framework for assessing the sustainability trade-offs of LLM-based chatbot design.

7 Limitation

This study is not without its limitations. First, our energy estimates rely on API end-to-end response time rather than direct server-side response time. This challenge is a common constraint in API-based LLM research, since proprietary services generally do not provide detailed server-side information needed for more transparent physical measurement. Recent work notes that API-served LLMs largely remain "black boxes" whose energy costs depend heavily on what providers choose to

disclose (Krupp et al., 2026). Therefore, our estimates should be interpreted as application-level approximations. Second, our experiments are limited to single question-response interactions. Future work could therefore expand the empirical scope of this framework by testing in sessions, with a larger set of LLMs and datasets. Last, the domain products in this experiment are linear RAG-based pipelines; future research should pivot toward "Agentic Workflow" (Yao et al., 2023; Ye et al., 2023; Wang et al., 2026) frameworks and MoE architecture (Shazeer et al., 2017) that offer synergistic pathways.

Acknowledgments

This work was supported by grants from the IKEA Foundation and William and Flora Hewlett Foundation to A.H. The authors acknowledge the contributions of Mariam Maher, Arboretica, in the algorithmic design of ChatNetZero. The development of ChatNDC was supported by an award from the Climateworks Foundation, and the University of North Carolina at Chapel Hill supported the development of ChatNetZero through a seed grant to A.H.

References

2026. Tiktoken. Original-date: 2022-12-01T23:22:11Z.

M. Adamska, D. Smirnova, H. Nasiri, Z. Yu, and P. Garraghan. 2026. Green Prompting: Characterizing Prompt-driven Energy Costs of LLM Inference. *arXiv*.

Ariyanti Sri and Suryanegara Muhammad. Trade-Off Between Energy Consumption and Three Configuration Parameters in Artificial Intelligence (AI) Training: Lessons for Environmental Policy.

Anja Belz and Ehud Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv preprint*. ArXiv:2305.05176 [cs].

Sophia Chen. 2025. How much energy will AI really consume? The good, the bad and the unknown. *Nature*, 639(8053):22–24. Bandiera_abtest: a Cg_type: News Feature Subject_term: Machine learning, Computer science, Sustainability, Energy.

Michael Copley. 2025. Data centers are booming. But there are big energy and environmental risks. *NPR*.

Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2025. The rising costs of training frontier AI models. *arXiv preprint*. ArXiv:2405.21015 [cs].

Maximilian Dauner and Gudrun Socher. 2025. Energy costs of communicating with AI. *Frontiers in Communication*, 10:1572947.

Cooper Elsworth, Keguo Huang, David Patterson, Ian Schneider, Robert Sedivy, Savannah Goodman, Ben Townsend, Parthasarathy Ranganathan, Jeff Dean, Amin Vahdat, Ben Gomes, and James Manyika. 2025. Measuring the environmental impact of delivering AI at Google Scale. *arXiv preprint*. ArXiv:2508.15734 [cs].

Brad Everman, Trevor Villwock, Dayuan Chen, Noe Soto, Oliver Zhang, and Ziliang Zong. 2023. Evaluating the carbon impact of large language models at the inference stage. In *2023 IEEE international performance, computing, and communications conference (IPCCC)*, pages 150–157. IEEE.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models. *arXiv preprint*. ArXiv:2309.14393 [cs].

Zhenxiao Fu, Fan Chen, Shan Zhou, Haitong Li, and Lei Jiang. 2025. LLMCO$_2$: Advancing Accurate Carbon Footprint Prediction for LLM Inferences. *ACM SIGEnergy Energy Informatics Review*, 5(2):63–68.

Google. Power usage effectiveness – Google Data Centers.

Giovanni Cocca Guardia, Gabriel Hermosilla Vigneau, Danny Espín-Sarzosa, and Bárbara Dumas Feris. 2024. Assessing the Energy Impact and Carbon Footprint of AI Model Training: A Case Study Using GPU Servers.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNetZero, an LLM-Chatbot to Demystify Climate Pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

Erik Johannes Husom, Arda Goknil, Lwin Khin Shar, and Sagar Sen. 2026. The Price of Prompting: Profiling Energy Use in Large Language Models Inference. *arXiv preprint*. ArXiv:2407.16893 [cs].

IEA. 2025. Energy demand from AI – Energy and AI – Analysis.

Intel. Intel Xeon Processor.

Nidhal Jegham, Marwan Abdelatti, Chan Young Koh, Lassad Elmoubarki, and Abdeltawab Hendawi. 2025. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. *arXiv preprint*. ArXiv:2505.09598 [cs].

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.

Lars Krupp, Daniel Geißler, Francisco M. Calatrava-Nicolas, Vishal Banwari, Paul Lukowicz, and Jakob Karolus. 2026. This Is Taking Too Long – Investigating Time as a Proxy for Energy Consumption of LLMs. *arXiv preprint*. ArXiv:2603.15699 [cs] version: 1.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint*. ArXiv:1910.09700 [cs].

Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2025. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv preprint*. ArXiv:2304.03271 [cs].

Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *The 2024 ACM Conference on Fairness Accountability and Transparency*, pages 85–99. ArXiv:2311.16863 [cs].

Diego Manya-Gutierrez, Alicia Bao, Jiamian He, Angel Hsu, and Ji (James) Zhang. 2026. Replication Data for: The Energy Footprint of LLM-based Environmental Analysis: LLMs and Domain Products.

Sam Meredith. 2023. A 'thirsty' generative AI boom poses a growing problem for Big Tech. Section: COP28.

Microsoft. Measuring energy and water efficiency for Microsoft Datacenters.

Sophia Nguyen, Beihao Zhou, Yi Ding, and Sihang Liu. 2024. Towards Sustainable Large Language Model Serving. *ACM SIGEnergy Energy Informatics Review*, 4(5):134–140.

James O'Donnell and Casey Crownhart. 2025. We did the math on AI's energy footprint. Here's the story you haven't heard. *MIT Technology Review*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to Route LLMs with Preference Data. *arXiv preprint*. ArXiv:2406.18665 [cs].

C. P. Ormell. 1974. Bloom's Taxonomy and the Objectives of Education. *Educational Research*, 17(1):3–18.

Soham Poddar, Paramita Koley, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models.

S. Podder, H. Date, and S. Murthy. 2026. Green prompt engineering for sustainable generative AI. *Environmental Science and Ecotechnology*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint*. ArXiv:1701.06538 [cs].

Samarth Sikand, Rohit Mehra, Priyavanshi Pathania, Nikhil Bamby, Vibhu Saujanya Sharma, Vikrant Kaulgud, Sanjay Podder, and Adam P. Burden. 2025. Breaking the ICE: Exploring promises and challenges of benchmarks for Inference Carbon & Energy estimation for LLMs. In *2025 IEEE/ACM 9th International Workshop on Green and Sustainable Software (GREENS)*, pages 55–59. ArXiv:2506.08727 [cs].

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv preprint*. ArXiv:2501.09136 [cs].

Net Zero Tracker. 2025. Net Zero Stocktake 2025. Technical report, NewClimate Institute, Oxford Net Zero, Energy and Climate Intelligence Unit, and Data-Driven EnviroLab.

Yuntao Wang, Shaolong Guo, Yanghe Pan, Zhou Su, Fahao Chen, Tom H. Luan, Peng Li, Jiawen Kang, and Dusit Niyato. 2026. Internet of Agents: Fundamentals, Applications, and Challenges. *IEEE Transactions on Cognitive Communications and Networking*, 12:4476–4501. ArXiv:2505.07176 [cs].

Hui Wu, Xiaoyang Wang, and Zhong Fan. 2025. Addressing the sustainable ai trilemma: a case study on llm agents and rag. *Preprint*, arXiv:2501.08262.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint*. ArXiv:2210.03629 [cs].

Yining Ye, Xin Cong, Shizuo Tian, Jiannan Cao, Hao Wang, Yujia Qin, Yaxi Lu, Heyang Yu, Huadong Wang, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. ProAgent: From Robotic Process Automation to Agentic Process Automation. *arXiv preprint*. ArXiv:2311.10751 [cs].

A Model output size

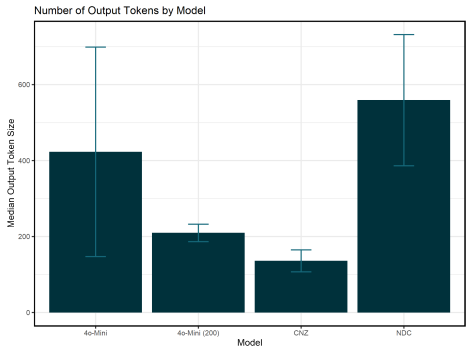


Figure A.1: Median answer token size by each model.

B Dataset representative questions

Question type	Example questions
Knowledge	"Has Marathon Petroleum made any climate pledges?"; "Does Foxconn have any climate strategy?"; "What is the total emissions reduction target of Malaysia?"; "Does Taiwan have a carbon trading scheme or policy under its latest NDC?"
Comprehension	"Interpret how ESG regulations influence business strategy and investor decisions"; "Explain the relationship between local air pollution and public health outcomes."
Application	"What are Scope 3 emissions and what should be included in a net-zero target?"; "If an entity doesn't control out-of-boundary emissions, how can it credibly set a net-zero target?"
Analysis	"Are Costco's or Walmart's net zero target 1.5 degree aligned?"; "How does Walmart's climate goals compare with other large retail stores?"; "How do England and Scotland NDCs compare to the UK's?"
Evaluation	"Is Apple's net zero target credible?"; "How is supermarket chain Jeronimo Martins doing on their commitment to reduce food waste by 50% by 2025?"
Creation	"Propose an ESG audit model that integrates social, environmental, and ethical performance indicators into one scorecard."; "Develop a set of innovative KPIs to measure corporate progress toward net-zero goals beyond traditional carbon metrics."

Table B.1: Taxonomy of example questions across different cognitive levels.

Can Multimodal Large Language Models Identify Thematic Maps in Scientific Articles?

Tomas Turek[1, 2], **Vincent Nguyen**[1], **Dhirendra Singh**[1], **Sarvnaz Karimi**[3], **Xiuzhen Zhang**[2]
[1]CSIRO (Australia), [2]RMIT University (Australia), [3]Monash University (Australia)
Correspondence: firstname.lastname@csiro.au, sarvnaz.karimi@monash.edu, xiuzhen.zhang@rmit.edu.au

Abstract

Information seeking systems have yet to access multimodal information about thematic maps in scientific literature. Thematic maps are images of geographical maps showing themes of information and are crucial for understanding the impact of climate on natural resources. Although Multimodal Large Language Models (MLLMs) are applied in different applications, their effectiveness in identifying thematic maps is not explored. We address this gap by evaluating some of the state-of-the-art MLLMs to (1) classify thematic maps using different visual and textual contexts; (2) explain their predictions and uncover their sources of bias using *attribute discovery*; and (3) assist humans with the annotation of thematic maps in the scientific literature. We release a new public dataset[1] and use it to show that MLLMs that were investigated in our study have room for improvement. We observe that MLLMs can be biased toward visual attributes of thematic maps (graphical indicators). The best classification result is an 80% strict F1 score by the Grok 4.3 model, which can fluctuate between 76% and 84% due to classification bias. Overall, we consistently find that the input of images with captions and examples of the correct answers yields substantial gains in task performance.

1 Introduction

Current multimodal tasks lack connection to the early work exploring the information in thematic maps (Tousidou and Manolopoulos, 2000; d'Onofrio and Pourabbas, 2001; Mauro and Ardissono, 2017). Understanding complex geographical context and user demands remains a challenge when searching for maps, even in structured sources of information (Gui et al., 2024). For example, it is difficult to classify themes in images and text metadata to help users find relevant maps from spatial databases (Yang et al., 2020; Wei et al., 2021). Emerging multimodal retrieval applications critically need methods to identify complex visual-textual elements in documents (Qiao et al., 2025).

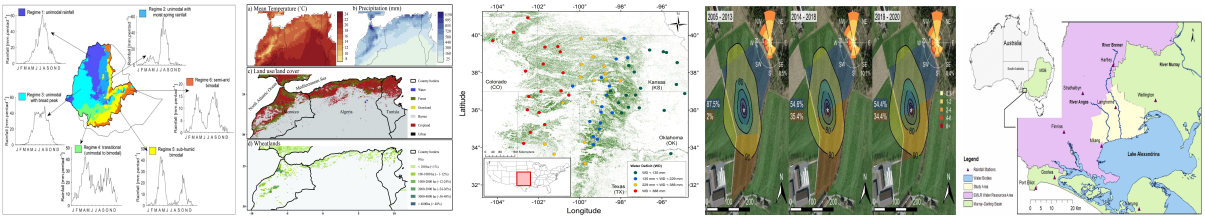
To illustrate, identifying climate adaptation strategies for a specific location requires climate data and knowledge of the scientific literature in text and figures. One type of such figures is thematic maps, which allow analysing the spatial distribution of thematic information in a specific region. For a system that retrieves relevant literature, such as a question-answering system for climate adaptation (Nguyen et al., 2025), identifying these maps in scientific articles could be a first step.

MLLMs are a key candidate for this task, but there is no study to investigate how effective they would be in detecting such figures in scientific articles. To bridge this gap, we investigate the classification of *thematic maps* in agricultural adaptation journals (examples shown in Figure 1). Thematic maps are spatial maps that illustrate particular themes in a landscape, such as precipitation in regional areas, and require specialised knowledge to interpret. We define a thematic map as an image that includes (i) a **base map** showing geographical features and (ii) a **thematic overlay** (Monmonier, 2001) showing measurable climatic parameters, such as land and sea surface temperatures. This is an essential theme of remote sensing information required for climate monitoring (WMO, 2025) that practitioners seek to better assess climate-related risks (Alonso-Robisco et al., 2026).

We evaluate closed- and open-source MLLMs on their ability to classify figures from scientific articles that contain thematic maps, using the image, its caption, and related discourse (text sentences that reference the figure) as input. Varying the input data demonstrates that the length and relevance of the textual context influence the performance of classifiers, e.g., combining shorter captions with longer scientific discourse degrades the output of

[1] https://github.com/tomtx/cp-thematic-maps

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 231–245
October 28, 2026 ©2026 Association for Computational Linguistics



(a) chorochromatic map (b) choropleth map (c) dot distribution map (d) isopleth map (e) general reference map

Figure 1: Examples of thematic map types in agricultural adaptation showing climatic parameters: (a) rainfall regimes, (b) precipitation and temperature, (c) water deficit, and (d) wind climatology. Example of other maps (e).

models. Analysing the visual attributes of map figures that lead to correct classifications reveals complex and spurious features, showing that models are biased toward directional, numerical, and physical representations on maps (graphical indicators) and struggle with colours and layouts of figures (gradients and subplots).

In this context, we advance the MLLM-assisted processing of multimodal research evidence in climate science with three contributions: (i) a novel classification task of thematic map identification in scientific articles with user instructions; (ii) a novel attribute discovery methodology to explain predictions of MLLMs in the classification of thematic maps; and (iii) a new data source of human-annotated figures with heterogeneous thematic maps of various types and locations in public, peer-reviewed scientific articles (CC BY-NC-SA 4.0). This assists downstream tasks in scientific domains that require multimodal evidence in scientific figures.

2 Related Work

Classification of Scientific Figures The closest work to our research is the ACL-FIG pipeline (Karishma et al., 2023) for a multi-class classification of images into standard types of figures, such as bar charts and maps, with 83% accuracy (vision transformer). They extract images from scientific articles, cluster and annotate image vectors into figure types, and link the figure metadata. Previous methods only label whether the image is a figure or table, such as DEEPFIGURES (Siegel et al., 2018). Our efforts advance ACL-FIG with a *multimodal classification* of scientific figures by (i) a classification of multimodal context of figures (images with texts instead of images only) and (ii) a fine-grained classification of map figures (thematic maps with more complex information instead of maps only).

Datasets with Map Figures Other works on benchmark datasets with map figures do not capture specialised data representations of different maps in scientific domains to enable multimodal classification. Benchmarks either contain unannotated or unimodal figures as the images are not labelled nor accompanied by captions or other relevant texts from their source. For example, SCIFIBENCH (Roberts et al., 2024) contains captioned figures, some of which are maps, but lacks annotated labels to understand whether the image is a box plot or a map and which article it originates from to contextualise the data. MAPQA (Chang et al., 2022) is the only dedicated dataset for thematic maps showing health indicators in choropleths. However, this is not a source of scientific figures, given the lack of captions. In addition, it is overly homogeneous and does not represent common types and spatial context of thematic maps, as the data are distributed over the same geographic region map.

Geospatial Tasks Images of thematic maps, which are published in scientific articles, are generally constructed by processing various remote sensing imagery (Foody, 2004; Schowengerdt, 2007). For example, geo-referenced aerial and satellite imagery are the geospatial data that are often used as base maps of thematic maps. MLLMs currently underperform in fine-grained object classification of multimodal geospatial data in the computer vision literature, e.g., below a 60% accuracy on GEOBENCH (Danish et al., 2025). This is due to their sensitivity to distribution shifts and prompt variations and their weak numerical, temporal, and spatial reasoning capabilities in complex geospatial scenes. However, such complex images of base maps do not contain thematic overlays captured by remote sensing technologies (such as climatic parameters) to understand the performance of MLLMs on thematic maps in geospatial tasks.

Visual Bias of Classifiers Previous efforts in classification tasks do not analyse visual attributes of scientific figures to understand spurious correlations between their labels and attributes. It is known that classifiers trained on biased images underperform on images that conflict with bias, e.g., misclassifying waterbirds (labels) on land backgrounds (attributes) as landbirds (Yenamandra et al., 2023). Most of the work is done on natural images with vision classifiers, some of which use language models to generate descriptions of images in order to better understand attributes in which classifiers fail (Ghosh et al., 2025).

Although MLLMs are generally not used as classifiers, other tasks report on their bias toward particular attributes in images. MLLMs over-rely on learnt spurious correlations in training data and then poorly generalise on out-of-domain data, e.g., developing a class-dependent spurious bias in object detection tasks (Hosseini et al., 2025). MLLMs are also prone to visual hallucinations, which is a misalignment between the responses generated and the visual evidence provided, in question answering tasks (Suri et al., 2025).

3 Methodology

Multimodal Classification We propose a *multimodal classification* task: given an image and text of a map figure, the task is to correctly identify a thematic map. We formulate the task as a multi-label classification of themes since a map figure can contain multiple plots of thematic maps, each with multiple themes. This is a problem of instance classification, where each instance may have more than one correct class. We expect that models already learnt during their training a mapping function between instances in a multimodal dataset and their classes to assign the correct class labels to a given instance. In other words, $f : ({x_v}^{(i)} \times {x_t}^{(i)}) \rightarrow y^{(i)}$, where $y \in \bigcup_{j=1}^{n} y_j$ represents the classes (themes), n is the number of classes, and x is the feature input vector of textual and visual modalities, t and v respectively, for the instance i (map figure).

We leverage MLLMs to classify the map figure into eight classes that pertain to thematic maps. MLLMs need to generate a multi-response answer to a multi-choice question from users. The question is defined by our domain experts as a command: *"Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none."*.

Attribute Discovery We aim to explain the predictions of MLLMs on our classification task by systematically analysing the visual attributes of map figures. In this context, a *visual attribute* refers to a distinct visual feature of the image, such as legends, grid lines, and displayed values. For this purpose, we define a *target attribute* as any visual feature under investigation to determine its impact on classifier performance. Since a single figure can simultaneously contain multiple visual attributes, these target attributes are not mutually exclusive and can overlap both within a single figure and across multiple figures. To isolate the specific effects of these target attributes, we categorise them into *favourable* and *unfavourable* attributes based on whether they correlate with an increase or decrease in classifier performance, respectively. Ultimately, we define a *biased* attribute as a favourable attribute that introduces the risk of shortcut learning due to spurious correlations between task-relevant labels and task-irrelevant attributes (Yenamandra et al., 2023).

With the above definitions, we propose an *attribute discovery* methodology leveraging MLLMs to explain the general classification performance of classifiers. To identify target attributes within figures, MLLMs serve as hypothesis generators and self-report on their bias through a three-step process by generating (i) detailed textual descriptions of images; (ii) hypotheses about target attributes (which influence classifier performance due to inherent data and model biases) using randomly selected samples of captions and generated descriptions; and (iii) labels to denote whether a given figure caption or image contains a target attribute. Subsequently, these target attributes are filtered and validated through a four-step process by selecting (i) attributes present in a significant sample size of figures; (ii) favourable and unfavourable attributes that are measured by changes in classifier performance when these attributes are present versus absent in the figures; (iii) biased attributes from the subset of favourable attributes, which introduce disproportionate increases in classifier performance when present, indicating the exploitation of a spurious correlation (shortcut learning); and (iv) a small subset of figures for manual inspection to verify the presence of bias in the ground truth.

We discover such a bias of proprietary MLLMs in visual attributes by modifying the LADDER methodology (Ghosh et al., 2025) to eliminate their specialised training of classifiers. Our training-free approach evaluates classifiers at inference time via MLLM-generated: (i) descriptions of images to provide context on visual attributes, which is common when contextualising figures (Medisetti et al., 2024); (ii) hypotheses about biased attributes from captions and generated descriptions of images in both correct and incorrect classifications (instead of LLM-generated hypotheses from captions only); and (iii) labels for each hypothesis in a figure.

4 Data Annotation

We curate our own multimodal dataset through human-model collaboration to extract, select, and annotate the figures. The curated dataset contains 1,208 annotated maps from 195 publicly available agricultural adaptation journals (CC BY 4.0) published between 1998 and 2024[2], each with an image, caption, discourse, DOI, annotated themes (ground truth answers to classification questions), generated description, and a list of generated attributes; see Table 1.

Extraction of Multimodal Figure Data Our domain experts review the literature for representative quality journals on agricultural adaptation in climate science and create a set of domain keywords to include or exclude articles based on titles and texts. Then, an automatic processing (i) identifies peer-reviewed adaptation journals; (ii) selects articles that are publicly available, contain figures, and match adaptation measures and strategies; and (iii) extracts image, caption, and discourse of each figure in the article.

Selection of Map Figures MLLM and vision models classify images of extracted scientific figures into map figures. The predictions are then aggregated via ensemble learning (hard voting) and verified on a small annotated sample by a domain expert. We use OpenAI o3 (our choice) and GPT-4o (the best classifier of aerial image objects in GEOBENCH) models by answering a binary classification question from users, which our domain experts define as: *"Does the figure show a map (a representation of the land or water surface area with outlined area boundaries)?"*. We train our own classifier by performing supervised fine-tuning of a vision transformer (ViT Base Patch16 224 IN21k) on map images from the ACL-FIG dataset. Performance in a random sample of 100 annotated figures is the F1 score of 93% by OpenAI o3 (90% precision, 97% recall), 92% by OpenAI GPT-4o (89% precision, 96% recall), and 71% by our fine-tuned vision transformer (75% precision, 68% recall).

Figure Type	Thematic Map	Other Map
Count	336	872
Size	861k bytes [37k, 4M]	733k bytes [17k, 6M]
Resolution	6.2M pixels [110k, 14M]	5.7M pixels [114k, 16M]
Caption	55 words [4, 271]	37 words [4, 199]
	350 characters [25, 2k]	236 characters [19, 1k]
Discourse	86 words [0, 589]	55 words [0, 311]
	505 characters [0, 4k]	331 characters [0, 2k]

Table 1: Summary of our curated dataset (averages in size and ranges of individual figure characteristics).

Annotation of Thematic Maps Our annotation guidelines direct two climate scientist annotators and one computational scientist facilitator to identify thematic maps from the machine-selected map figures by (i) discussing the task and modifying its setup to ensure the annotation instructions are sound [30 minutes]; (ii) individually annotating a randomly selected 30 samples, evaluating the annotations to check for substantial agreement, and discussing the challenges together and adjusting the task [2 hours]; and (iii) individually annotating the full dataset [4 weeks at own pace]. Given the image with caption of a map figure and two examples of correct answers, annotators determine whether it is a thematic map by answering the following question formulated by our domain experts as: *"Does the figure show a climatic parameter, such as atmospheric pressure, evaporation, humidity, precipitation, radiation, temperature, or wind?"*. The facilitator resolves the ties and skipped annotations and labels the individual themes (climatic parameters) of each annotated thematic map. The result is a near-perfect inter-rater agreement of domain experts (91.87% Cohen's kappa coefficient).

The curated dataset can be generalised in other scientific domains across its annotated themes of map figures at various locations. Our pipeline is reproducible with other scientific articles given the reduced workload for human annotation by the semi-automated extraction and selection of figures. We use models to automatically select map figures since manual selection is not feasible, even though models may introduce a selection bias when select-

[2]XML articles with JPEG figures from Elsevier. https://dev.elsevier.com

ing a subset of figures with maps for human annotation. However, models rarely miss maps when aggregating their predictions, omitting only 46 from 16,025 figures. Although the scale of data is relatively small, this is currently the largest collection of map figures compared to other benchmarks with scientific figures (only 53 and 100 maps are in SCIFIBENCH and ACL-FIG datasets, respectively).

5 Experiments

The experimental setup for the classification task evaluates MLLMs (configured as in Table 5) in zero- and few-shot learning settings, both with and without Chain-of-Thought (CoT) prompting. To determine which context in the scientific article yields the best classification performance, we investigate two input combinations: the image paired with its caption, and the image paired with both its caption and discourse. Furthermore, to understand the suitability of MLLMs for assisting humans with the identification of research evidence, we align one of the prompts with the annotation instructions given to our domain experts. Specifically, we supply the models with an image, its corresponding caption, and two exemplar answers without CoT.

For few-shot learning, prompts include a positive and negative exemplar. We evaluate in-distribution capabilities by sourcing them directly from our curated dataset to provide models with high-fidelity, domain-specific demonstrations of map figures. The exemplars are selected via stratified random sampling from a pool of 10 expert-curated exemplars with an equal class split, excluding the evaluation instance to prevent data leakage.

For CoT prompting, we experiment with both implicit and explicit commands. The implicit command appends the *"Let's think step by step."* instruction proposed by Kojima et al., 2022 (see Figures 4 and 7). Conversely, the explicit command replicates the reasoning process of our domain experts via a step-by-step procedural instruction. We handcrafted a sequence of reasoning steps required to complete this task, focusing on both the textual and visual modalities of the figures (see Figure 5). For few-shot learning, our exemplars are augmented with the reasoning steps taken following the approach of Wei et al., 2022 (see Figure 8).

We measure the effectiveness of models across a set of classes using F1 scores for fine-grained type problems (Ling and Weld, 2021). The F1 score is computed from precision and recall metrics, which we calculate using strict, loose macro, and loose micro variations. The strict version defines both metrics as the unweighted mean of the exact match between the predicted and labelled set of classes for each instance. The loose macro version calculates the unweighted mean of each metric across classes for each instance. Finally, the loose micro version sums the dividends and divisors across classes for each instance before computing the final quotient.

For the attribute discovery experiments, we explain the predictions made under the best-performing configuration to better understand classifier performance. We prompt each MLLM in a zero-shot setting to generate concise, three-sentence descriptions of the figure images, as well as hypotheses about the target attributes (including attribute names and examples within map figures). The inputs for hypothesis generation consist of randomly selected captions paired with the generated descriptions of figures from both correct and incorrect classifications (100 samples per category). Following established protocols (Yenamandra et al., 2023; Hosseini et al., 2025; Ghosh et al., 2025), we then analyse spurious features in the images by comparing a set of unmodified images containing the target attributes against a distinct set of unmodified images lacking those attributes.

To filter these target attributes for analysis, we establish thresholds for both model performance and sample size; this ensures the discovered attributes represent distinct groups of common visual attributes on maps that demonstrably influence classifier performance. To set the model performance threshold, we select the five attributes that yield the largest absolute changes in classifier effectiveness (measured by the strict F1 score). To set the sample size threshold, we select attributes present in more than 10% of the figures in the classification dataset. Relaxing these filtering criteria would increase the likelihood of sampling errors and surface less representative attributes (see Table 6). For example, the 'blue-red contour' attribute appears in only 4% of the maps and represents a minor subset of the broader 'blue-red gradient' attribute, which captures 21% of the maps.

6 Results

The results shown in Table 2 indicate that the highest performance in thematic map identification is an 80% strict F1 score by the Grok 4.3 model. This result was achieved using an image with caption input

Model	Source	CoT	Zero-Shot Learning						Few-Shot Learning					
			Image + Caption			Image + Caption + Discourse			Image + Caption			Image + Caption + Discourse		
			Strict	*Macro*	*Micro*	*Strict*	*Macro*	*Micro*	*Strict*	*Macro*	*Micro*	*Strict*	*Macro*	*Micro*
Claude Opus 4.6	closed		60.0	66.7	61.7	35.7	42.6	39.9	69.5	75.2	71.8	62.5	69.1	65.3
Gemini 3.5	closed		72.4	78.0	73.3	60.5	67.1	61.5	76.3	81.1	76.0	71.0	76.6	71.7
GPT-5.5	closed		74.5	79.3	74.6	66.8	73.1	68.1	76.3	80.8	75.9	70.0	75.4	70.0
Grok 4.3	closed		70.1	75.2	65.3	64.4	70.8	61.2	79.9	82.3	75.7	75.5	79.3	72.6
Kimi K2.5	open	$\mathcal{X}$	39.0	46.5	42.0	28.2	36.4	32.8	n/a	n/a	n/a	n/a	n/a	n/a
Llama 4 Maverick	open		44.8	51.3	50.4	32.9	39.3	41.0	71.0	76.9	72.3	54.1	60.6	56.3
Gemma 4	open		74.4	80.2	74.8	70.0	75.7	70.6	76.3	81.6	76.4	68.0	72.6	67.8
DeepSeek V4	open		67.3	74.1	67.8	59.0	66.6	60.8	74.1	80.9	75.2	76.0	82.2	75.5
Mistral Large 3	open		9.7	15.5	16.5	6.0	12.7	14.1	n/a	n/a	n/a	n/a	n/a	n/a
Qwen3-VL	open		63.8	70.0	64.7	39.6	46.5	46.5	n/a	n/a	n/a	n/a	n/a	n/a
Claude Opus 4.6	closed		56.3	64.2	56.8	45.9	54.1	48.9	69.9	76.0	72.1	63.7	70.4	66.3
Gemini 3.5	closed		72.3	78.0	73.0	60.5	67.3	61.6	77.2	82.1	**77.2**	72.1	77.7	72.4
GPT-5.5	closed		73.3	78.4	73.9	67.5	73.2	68.2	76.1	80.5	75.6	70.4	75.6	70.4
Grok 4.3	closed		69.7	75.2	65.3	64.0	70.1	59.8	**80.0**	**82.4**	75.6	75.1	79.0	72.3
Kimi K2.5	open	$\mathcal{I}$	34.2	41.4	35.4	30.0	38.6	34.0	n/a	n/a	n/a	n/a	n/a	n/a
Llama 4 Maverick	open		40.5	47.2	47.4	29.8	36.1	37.9	69.9	76.0	71.6	49.8	56.1	52.4
Gemma 4	open		72.3	78.1	73.0	68.7	74.6	69.8	75.1	80.5	76.1	67.2	72.0	67.3
DeepSeek V4	open		66.8	74.0	67.5	58.8	66.2	58.7	71.5	79.0	73.0	74.4	81.1	74.2
Mistral Large 3	open		8.8	15.0	16.2	6.9	13.2	14.6	n/a	n/a	n/a	n/a	n/a	n/a
Qwen3-VL	open		67.1	72.9	66.5	47.1	54.0	51.6	n/a	n/a	n/a	n/a	n/a	n/a
Claude Opus 4.6	closed		31.5	33.4	28.2	21.5	22.8	18.0	70.9	76.6	72.2	66.3	72.4	67.0
Gemini 3.5	closed		70.9	76.5	70.9	61.9	68.8	63.3	76.7	81.5	76.6	70.3	75.9	70.8
GPT-5.5	closed		71.4	76.2	70.6	64.0	70.2	65.2	74.2	78.6	73.4	68.0	73.7	68.3
Grok 4.3	closed		71.2	76.0	67.8	61.8	68.4	59.0	78.1	80.9	74.4	72.0	76.2	69.2
Kimi K2.5	open	$\mathcal{E}$	38.8	45.5	39.8	42.1	49.6	42.7	n/a	n/a	n/a	n/a	n/a	n/a
Llama 4 Maverick	open		29.1	35.5	38.6	18.9	25.6	30.0	62.3	68.5	65.7	28.1	34.9	34.5
Gemma 4	open		67.7	74.0	68.8	63.0	69.5	65.1	56.6	61.2	59.4	10.7	12.3	13.4
DeepSeek V4	open		49.3	56.9	49.3	43.0	51.3	44.6	65.1	72.6	66.1	64.2	72.0	64.8
Mistral Large 3	open		17.9	23.3	22.9	21.7	27.0	25.2	n/a	n/a	n/a	n/a	n/a	n/a
Qwen3-VL	open		25.2	32.0	41.4	26.9	34.4	40.7	n/a	n/a	n/a	n/a	n/a	n/a

Table 2: Classification of themes in map figures using input context of image with caption or image with caption and discourse via zero- and few-shot learning without ($\mathcal{X}$) or with Implicit ($\mathcal{I}$) or Explicit ($\mathcal{E}$) Chain-of-Thought (CoT) prompting. Evaluation metrics are strict, loose macro, and loose micro F1 scores (n/a due to model limits).

under few-shot learning with implicit CoT prompting, which we found to be the best configuration for multi-label classification across MLLMs supporting multimodal few-shot exemplars. However, the performance does not align well with human judgement as classifiers are generally unable to follow human-prescribed reasoning steps (explicit CoT) and only self-generated reasoning pathways (implicit CoT) improve their performance.

MLLMs struggle to classify map figures with blue-red colours (56% by DeepSeek V4) and multiple plots (64% by Llama 4 Maverick) as per Table 3. A further analysis of the classifications reveals that the performance of models also drops when removing the figures with visual attributes on which the models are biased. Such attributes are scale bars, north arrows, and river outlines that represent numerical trends, spatial directions, and physical waterway flows on maps, respectively.

Q1: Does relevant text in scientific articles improve the identification of thematic map figures by MLLMs? Classifiers perform the thematic map identification task best without a discourse in their input, relying solely on images paired with their captions. This text context generally does not improve performance and only a few MLLMs benefit from its inclusion. Specifically, Kimi K2.5, Mistral Large 3, and Qwen3-VL yield minor gains of up to 3 points of strict F1 scores in zero-shot learning with explicit CoT prompting as per Table 2. This discourse-related degradation is not due to context window constraints, as it persists under the reduced context of zero-shot learning, where classifiers processed complete inputs.

Our hypothesis is that scientific knowledge in discourse becomes too complex for MLLMs to interpret accurately when linked to the already dense context within the image of a map figure. For example, only domain experts may understand whether the scientific discussion about adaptation strategies in discourse is relevant to particular themes of climatic parameters in a map figure, since such themes are sometimes not explicitly mentioned in any input of image, caption, or discourse. MLLMs may lack this implicit domain knowledge to bridge such modal gaps. In addition, an image with a caption generally contains richer and more closely related information about a thematic map figure than any other relevant text in an article (discourse).

Our work builds upon ablation studies examining the input context of figures provided to multimodal systems, demonstrating that combining images with captions outperforms using images alone for question answering over scientific articles (Pra-

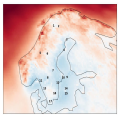
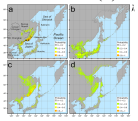
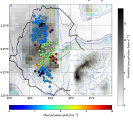
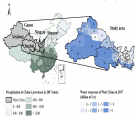
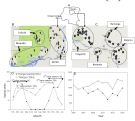
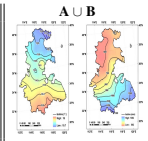
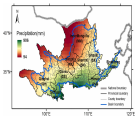
Model	Source	Attribute										Attribute Union			
		Blue-Red Gradient (A)		Multi-Panel (B)		Scale Bar (C)		North Arrow (D)		River Outline (E)		A ∪ B		C ∪ D ∪ E	
															
		137 thematic maps 118 other maps		297 thematic maps 524 other maps		149 thematic maps 541 other maps		146 thematic maps 496 other maps		39 thematic maps 192 other maps		310 thematic maps 547 other maps		178 thematic maps 624 other maps	
		Present	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*
Claude Opus 4.6	closed	57.6	73.1	64.4	81.4	75.2	62.7	74.6	64.5	73.2	69.1	64.6	82.6	74.1	61.6
Gemini 3.5	closed	67.5	79.9	72.8	86.6	80.3	73.2	81.3	72.6	80.5	76.5	72.8	88.0	80.0	71.1
GPT-5.5	closed	68.6	78.1	72.2	84.2	78.4	73.0	78.7	73.1	78.8	75.4	72.2	85.5	78.2	71.9
Grok 4.3	closed	70.6	82.5	75.5	89.4	81.9	76.8	82.9	76.7	84.0	79.0	75.5	90.9	82.2	75.6
Llama 4 Maverick	open	56.2	73.6	63.7	82.9	75.5	62.4	76.5	62.4	76.2	68.4	63.9	84.3	74.8	60.1
Gemma 4	open	63.9	78.1	70.5	84.8	79.9	68.7	80.4	69.1	81.0	73.7	70.6	86.0	79.8	65.8
DeepSeek V4	open	56.1	75.7	64.8	85.8	76.7	64.7	77.3	65.0	74.9	70.7	65.3	86.6	76.1	62.6
influence on performance		unfavourable		unfavourable		favourable		favourable		favourable		unfavourable		favourable	

Table 3: Explanation of classification performance showing discovered attributes when present or absent in map figures, including maps with multiple sets of attributes (union of maps) and their influence on performance. Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

manick et al., 2024). Collecting a caption and reference text about a figure in an article is standard practice in processing scientific figures (Sundar et al., 2024), but existing studies analysing fine-grained citation contexts overlook figure citations (Jantsch et al., 2025). Addressing this gap is essential to better contextualise the associated scientific discourse within articles. To avoid irrelevant context in our collected discourse, we follow the established protocol of extracting only the citing sentences rather than the broader surrounding text, which may include unrelated details about the figures.

Q2: Can examples of map figures with correct answers improve the classification performance of MLLMs? Performance of MLLMs improves when given the same instructions as annotators, i.e., few-shot examples of images with captions and answers to classification questions (compared to zero-shot). This is consistent with human requirements, as our domain experts need examples of answers to correctly perform the annotation task. Few-shot learning provides a maximum increase of 45 strict F1 points, achieved by Claude Opus 4.6 under explicit CoT prompting using images with captions and discourse (Table 2).

The ability of MLLMs to process multiple of multimodal examples and with high resolution and large size of images is limited. Models struggle with an increased volume of complex information in few-shot learning and are generally unable to process more than two examples due to their input context limits for both textual and visual modalities. Some models can process only one map image in their input, provided that it is small in size and low in resolution. For example, Kimi K2.5, Mistral Large 3, and Qwen3-VL cannot process inputs for our few-shot learning experiments and Claude Opus 4.6 limits the image size to 5MB. However, we did not reduce the size of the images as this reduced the classification performance of MLLMs.

Interestingly, the performance of MLLMs in interpreting figures may decrease when few-shot examples are included in the prompts (Roberts et al., 2024). We have only noticed such behaviour for the Gemma 4 model and mainly under its input of image with caption and discourse when going from zero-shot to few-shot learning. Fine-tuning and CoT prompting can mitigate some of the challenges of MLLMs in processing multimodal data for in-context learning (Zeng et al., 2024).

Q3: Are MLLM classifiers biased toward visual attributes of map figures? The goal is to understand visual attributes of map figures that lead to reliable classifications in practice. Therefore, we analysed attributes that influence results of MLLMs and validated whether these classifiers are biased toward such attributes. Table 3 shows an analysis of discovered attributes that are shared across MLLMs (self-reported by each model) for the best performing configuration of few-shot learning with implicit CoT prompting on images with captions.

We observed that the best classifier (Grok 4.3) correctly identified thematic maps when given map figures that show scale bars, north arrows, and river outlines. These attributes appear in about half of the thematic maps (178 out of 336). However, classification performance drops by nearly 7 points on the strict F1 score for figures without such visual attributes (representing the remaining half of the thematic maps). Our domain experts verified

Model	Source	Theme																				
		Atmospheric Pressure 10 map figures [1 with 1 theme, 4 with 2 themes, 4 with 3 themes, 1 with 4 themes]			**Evaporation** 118 map figures [13 with 1 theme, 22 with 2 themes, 37 with 3 themes, 46 with 4 themes]			**Humidity** 54 map figures [0 with 1 theme, 3 with 2 themes, 6 with 3 themes, 45 with 4 themes]			**Precipitation** 251 map figures [67 with 1 theme, 95 with 2 themes, 39 with 3 themes, 50 with 4 themes]			**Radiation** 36 map figures [8 with 1 theme, 8 with 2 themes, 14 with 3 themes, 6 with 4 themes]			**Temperature** 205 map figures [32 with 1 theme, 75 with 2 themes, 48 with 3 themes, 50 with 4 themes]			**Wind** 16 map figures [11 with 1 theme, 1 with 2 themes, 2 with 3 themes, 2 with 4 themes]		
		P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Claude Opus 4.6	closed	**83.3**	**50.0**	**62.5**	54.9	52.5	53.7	21.4	16.7	18.8	60.1	**86.5**	**70.9**	41.4	**66.7**	51.1	57.3	72.7	**64.1**	36.6	**93.8**	52.6
Gemini 3.5	closed	**83.3**	**50.0**	**62.5**	73.1	41.5	53.0	31.0	**16.7**	**21.7**	68.7	73.3	**70.9**	**55.9**	52.8	54.3	**66.8**	61.0	63.8	41.7	**93.8**	57.7
GPT-5.5	closed	**83.3**	**50.0**	**62.5**	65.4	44.9	53.3	18.2	11.1	13.8	66.8	73.7	70.1	55.3	58.3	**56.8**	66.5	56.1	60.8	41.7	**93.8**	57.7
Grok 4.3	closed	71.4	50.0	58.8	**69.4**	36.4	47.8	**42.9**	11.1	17.6	**69.3**	56.6	62.3	54.3	52.8	53.5	65.9	54.6	59.7	**50.0**	**93.8**	**65.2**
Llama 4 Maverick	open	**83.3**	**50.0**	**62.5**	63.2	36.4	46.2	27.8	9.3	13.9	61.7	75.7	68.0	48.6	50.0	49.3	51.3	75.6	61.1	42.9	**93.8**	58.8
Gemma 4	open	**83.3**	**50.0**	**62.5**	59.2	**60.2**	**59.7**	38.9	13.0	19.4	63.4	80.1	70.8	55.6	55.6	55.6	63.4	62.4	62.9	48.4	**93.8**	63.8
DeepSeek V4	open	66.7	20.0	30.8	68.6	29.7	41.4	13.6	5.6	7.9	59.0	83.7	69.2	40.5	47.2	43.6	48.7	**88.8**	62.9	40.0	75.0	52.2

Table 4: Analysis of binary classification errors per individual themes in map figures (each map can have multiple themes). Evaluation of best performing configuration via Precision (P), Recall (R), and F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

that these attributes are irrelevant to determining whether a map is a thematic map, since they do not relate to climatic parameter themes, and the model is biased toward such attributes of map figures.

Our analysis identifies a distinct model bias induced by a visual shortcut, in which MLLMs exploit spurious, non-causal associations between the class of thematic maps and the broader figure category of maps. Selection bias in classification datasets can amplify this observation when models omit difficult samples and retain only easy examples during data selection. Our curated dataset overcomes this issue by using a high-accuracy model ensemble to select a comprehensive collection of maps from the initial set of extracted figures. Although we verified a subset of the MLLM-generated descriptions for quality, we did not fully assess the faithfulness of these descriptions to the visual content of maps, nor did we measure how accurately MLLMs discover visual attributes or their robustness to prompt variance.

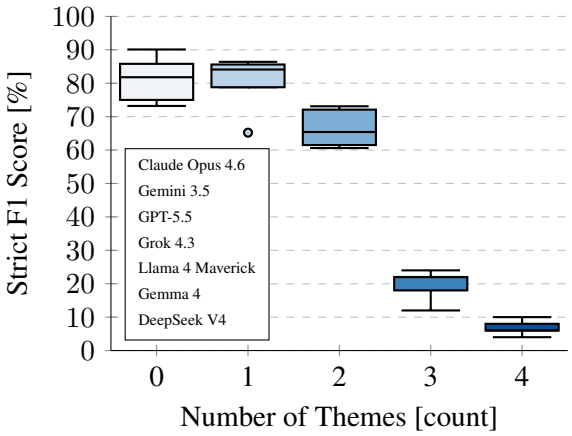


Figure 2: Analysis of multi-label classification errors per number of themes in map figures (0 themes in 872 maps, 1 in 132, 2 in 104, 3 in 50, and 4 in 50). Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

7 Error Analysis

The error analysis reveals challenges for MLLMs in understanding labels of climatic parameters and classifying multiple labels. We examined the results of the best performing configuration, which is few-shot learning with implicit CoT on the input context of images with captions. The classification performance per individual themes is shown in Table 4 and per number of themes in Figure 2.

The analysis of binary classification errors shows that the models cannot classify certain climatic parameters. The best F1 score is only 22% on thematic maps showing a single climatic parameter of humidity (see Table 4). For example, MLLMs misclassify 'precipitable water' data in thematic maps as 'precipitation', which domain experts classify as 'humidity' and 'atmospheric pressure' as these are the climatic parameters used to calculate the liquid water depth from condensed water vapour in the atmosphere. Our observation is that MLLMs do not correctly link climate science terminology explicitly stated in a caption or image of a map figure to climatic parameters. Other example is the 'RCP' term, which denotes Representative Concentration Pathways by amount of radiative forcing in near future, but models do not link it to (implicitly stated) 'radiation' climatic parameter and misclassify this as a 'none' theme. This demonstrates that MLLMs do not understand climatic parameters well, most likely due to the lack of data and instructions focused on numerical reasoning with climatic parameters during their training.

The analysis of multi-label classification errors reveals a performance decline with an increasing number of themes in the map figures. As illustrated in Figure 2, the evaluated MLLMs demonstrate a

strong baseline capability for single-label classification (zero or one theme), achieving median strict F1 scores over 80% for individual labels. However, performance decreases as task complexity increases, dropping to a median of approximately 65% for two labels, 20% for three labels, and under 10% for four labels. This 45 percentage point drop between two and three labels indicates that multi-label classification of multimodal data is a challenging task for MLLMs.

8 Conclusion

We evaluated whether MLLMs can identify thematic maps in scientific literature. This is a crucial step towards designing an information seeking system that can access different sources of multimodal data, such as climate data and text and figures from scientific articles, in answering complex questions on climate adaptation. We show that off-the-shelf MLLMs do not perform well on the classification of thematic maps found in scientific figures of maps. This poses risks as MLLMs are increasingly adopted as multi-purpose models in high-stakes applications, such as AI systems advising on climate adaptation. We find that classification performance is best when MLLMs are supplied with figures with related captions and human-verified examples of good classifications. Surprisingly, any additional context, such as a sentence referring to the respective figure, leads to worse performance. Further research is required to improve MLLM classification of multimodal scientific figures, investigate their ability to understand complex visual-textual context, such as one presented in thematic maps, and control and mitigate their classification bias.

Limitations

The task formulation is limited to classification to establish the baseline performance of MLLMs in identifying thematic maps. Another practical evaluation for a downstream use of MLLMs on such a dataset of thematic maps should follow. For example, question-answering for a deeper understanding of content in map figures.

The scope of our work is limited to maps in the practice of climate adaptation. Although map figures in the curated dataset contain other themes, such as land use and agricultural commodities, only climatic parameters are annotated. Collecting articles from various domains and annotating their map figures with themes would clarify the generalisability of thematic maps across different themes. For example, future work can investigate whether MLLMs perform similarly when classifying climatic parameters versus healthcare indicators.

Ethical Considerations

Only publicly available data were used in this research, and no additional data were collected. The authors obtained ethics approval from CSIRO for user studies with human experts, including their annotations (Climate Services for Agriculture, Phase 2, 131/24).

Acknowledgements

The authors gratefully acknowledge Willow Hallgren, Ashley Harkin, Vihan Weeraratne, and Mahesh Prakash for their agricultural and climate science expertise and their valuable contributions to the think-aloud studies. Tomas Turek is supported by the CSIRO PhD scholarship.

References

Andres Alonso-Robisco, Jose Manuel Carbo, Emily Kormanyos, and Elena Triebskorn. 2026. Houston, we have a problem: Can satellite information bridge the climate-related data gap? *Latin American Journal of Central Banking (LAJCB)*, 7(1):100173.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. MapQA: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation (TRL) Workshop*, New Orleans, Louisiana, United States.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2025. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7132–7142, Honolulu, Hawaii, United States.

Alberto d'Onofrio and Elaheh Pourabbas. 2001. Formalization of temporal thematic map contents. In *Proceedings of the 9th ACM International Symposium on Advances in Geographic Information Systems*, GIS '01, page 15–20, New York City, New York, United States. Association for Computing Machinery (ACM).

Giles M. Foody. 2004. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, 70(5):627–633.

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B Poynton, Shyam

Visweswaran, and Kayhan Batmanghelich. 2025. LADDER: Language-driven slice discovery and error rectification in vision classifiers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22935–22970, Vienna, Austria. Association for Computational Linguistics (ACL).

Zhipeng Gui, Xinjie Liu, Anqi Zhao, Yuhan Jiang, Zhipeng Ling, Xiaohui Hu, Fa Li, Zelong Yang, Huayi Wu, and Shuangming Zhao. 2024. Map retrieval intention recognition based on relevance feedback and geographic semantic guidance: For better understanding user retrieval demands. *Information Processing & Management*, 61(4):103767.

Parsa Hosseini, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. 2025. SpurLens: Spurious correlations in multimodal LLMs. In *Workshop on Spurious Correlation and Shortcut Learning (SCSL): Foundations and Solutions*, Singapore. The Thirteenth International Conference on Learning Representations (ICLR).

Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, and Young-Kyoon Suh. 2025. FineCite: A novel approach for fine-grained citation context analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24525–24542, Vienna, Austria. Association for Computational Linguistics (ACL).

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. ACL-Fig: A dataset for scientific figure classification. In *Proceedings of the Workshop on Scientific Document Understanding (SDU) co-located with 37th Conference on Artificial Inteligence (AAAI)*, volume 3656, Washington, DC, United States. CEUR-WS.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213. Curran Associates, Inc.

Xiao Ling and Daniel Weld. 2021. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI Technical Track: AI and the Web)*, volume 26, page 94–100, Toronto, Ontario, Canada. The Association for the Advancement of Artificial Intelligence (AAAI) Press.

Noemi Mauro and Liliana Ardissono. 2017. Thematic maps for geographical information search. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, page 337–342, New York City, New York, United States. Association for Computing Machinery (ACM).

Gowtham Medisetti, Zacchaeus Compson, Heng Fan, Huaxiao Yang, and Yunhe Feng. 2024. LitAI: Enhancing multimodal literature understanding and mining with generative AI. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 471–476, San Jose, California, United States. Institute for Electrical and Electronics Engineers (IEEE).

Mark Monmonier. 2001. Thematic maps in geography. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 15636–15641. Pergamon, Oxford, United Kingdom.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, and Mahesh Prakash. 2025. Question answering in climate adaptation for agriculture: Model development and evaluation with expert feedback. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7045–7075, Vienna, Austria. Association for Computational Linguistics (ACL).

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 118807–118833, Vancouver, Canada. Curran Associates, Inc.

Jingfen Qiao, Jia-Huei Ju, Xinyu Ma, Evangelos Kanoulas, and Andrew Yates. 2025. Reproducibility, replicability, and insights into visual document retrieval with late interaction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3335–3345, New York City, New York, United States. Association for Computing Machinery (ACM).

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. SciFIBench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 18695–18728, Vancouver, Canada. Curran Associates, Inc.

Robert A. Schowengerdt. 2007. Chapter 9 - thematic classification. In *Remote Sensing (Third Edition)*, third edition edition, pages 387–XXXIII. Academic Press, Burlington, Massachusetts, United States.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 223–232, New York City, New York, United States. Association for Computing Machinery (ACM).

Anirudh Sundar, Jin Xu, William Gay, Christopher Richardson, and Larry Heck. 2024. cPAPERS: A dataset of situated and multimodal interactive conversations in scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 66283–66304, Vancouver, Canada. Curran Associates, Inc.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. 2025.

ChartLens: Fine-grained visual attribution in charts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22447–22462, Vienna, Austria. Association for Computational Linguistics (ACL).

Eleni Tousidou and Yannis Manolopoulos. 2000. A performance comparison of quadtree-based access methods for thematic maps. In *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1*, SAC '00, page 381–388, New York City, New York, United States. Association for Computing Machinery (ACM).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zhengyang Wei, Zhipeng Gui, Min Zhang, Zelong Yang, Yuao Mei, Huayi Wu, Hongbo Liu, and Jing Yu. 2021. Text GCN-SW-KNN: a novel collaborative training multi-label classification method for WMS application themes by considering geographic semantics. *Big Earth Data*, 5(1):66–89.

WMO. 2025. The 2022 GCOS ECVs Requirements. Technical Report Global Climate Observing System (GCOS) Report-245, World Meteorological Organization (WMO), Geneva, Switzerland.

Zelong Yang, Zhipeng Gui, Huayi Wu, and Wenwen Li. 2020. A latent feature-based multimodality fusion method for theme classification on web map service. *Institute for Electrical and Electronics Engineers (IEEE) Access*, 8:25299–25309.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. 2023. FACTS: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4794–4804, Paris, France. Institute for Electrical and Electronics Engineers (IEEE).

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs perform text-to-image in-context learning? In *First Conference on Language Modeling (COLM)*, Philadelphia, Pennsylvania, United States.

A Appendix

This section details the model configurations, prompts, and additional evaluation results. The aim is to support the reproducibility of our empirical findings on thematic map identification within map figures, which are multimodal scientific figures sourced from scientific articles.

Model	Source	Provider	Release	Size	Context Window	Configuration
Claude Opus 4.6 V1	closed	Anthropic (United States)	5 February 2026	undisclosed	1M tokens	reasoning effort: medium retry limit: 3 temperature: 0
Gemini 3.5 Flash	closed	Google DeepMind (United Kingdom)	19 May 2026	undisclosed	1M tokens	reasoning effort: medium retry limit: 3 temperature: 0
GPT-5.5	closed	OpenAI (United States)	23 April 2026	undisclosed	1M tokens	reasoning effort: medium retry limit: 3
Grok 4.3	closed	SpaceXAI (United States)	17 April 2026	undisclosed	1M tokens	reasoning effort: medium retry limit: 3 temperature: 0
Kimi K2.5	open	Moonshot AI (China)	27 January 2026	1T parameters (32b active)	256k tokens	retry limit: 3
Llama 4 Maverick 17b Instruct V1	open	Meta AI (United States)	25 May 2025	400b parameters (17b active)	1M tokens	retry limit: 3 temperature: 0
Gemma 4 31B IT	open	Google DeepMind (United Kingdom)	2 April 2026	31b parameters (31b active)	256k tokens	reasoning effort: medium retry limit: 3 temperature: 0
DeepSeek V4 Flash	open	DeepSeek (China)	31 June 2026	284b parameters (13b active)	1M tokens	reasoning effort: medium retry limit: 3 temperature: 0
Mistral Large 3 675b Instruct	open	Mistal AI (France)	2 December 2025	675b parameters (41b active)	256k tokens	retry limit: 3 temperature: 0
Qwen3 VL 235B A22B	open	Alibaba Cloud (China)	23 September 2025	235b parameters (22b active)	256k tokens	retry limit: 3 temperature: 0

Table 5: Overview of benchmarked models on our multimodal classification dataset, including their configurations.

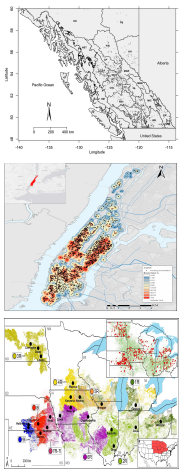
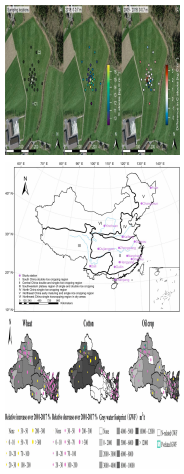
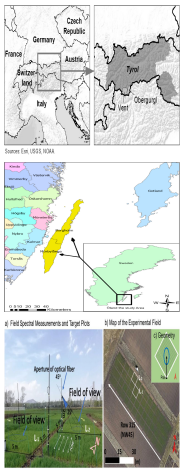
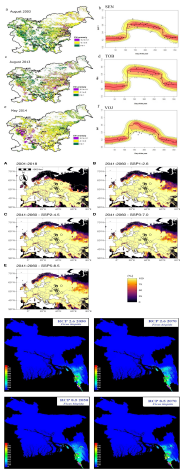
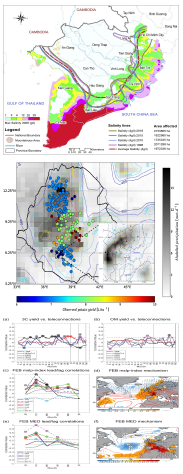
Model	Source	Attribute									
		Black Dot		Pink Dot		Flow Arrow		Year Label		Blue-Red Contour	
											
		69 thematic maps 150 other maps		5 thematic maps 29 other maps		20 thematic maps 104 other maps		59 thematic maps 85 other maps		16 thematic maps 33 other maps	
		Present	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*	*Present*	*Absent*
Claude Opus 4.6	closed	72.6	69.3	67.7	69.9	70.2	69.8	68.8	70.0	71.4	69.8
Gemini 3.5	closed	76.7	77.4	79.4	77.2	79.0	77.0	74.3	77.6	77.6	77.2
GPT-5.5	closed	77.2	75.8	70.6	76.2	78.2	75.8	73.6	76.4	77.6	76.0
Grok 4.3	closed	76.7	80.7	82.4	79.9	82.3	79.7	77.8	80.3	81.6	79.9
Llama 4 Maverick	open	73.1	69.2	73.5	69.8	72.6	69.6	66.7	70.3	73.5	69.7
Gemma 4	open	76.3	74.8	73.5	75.1	77.4	74.8	72.9	75.4	78.0	75.0
DeepSeek V4	open	75.8	70.6	73.5	71.5	73.4	71.3	68.1	72.0	69.4	71.6

Table 6: Explanation of classification performance showing discovered attributes when present or absent in map figures, specifically attributes that yield the lowest absolute changes in classifier effectiveness. Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

Classification Prompt: Zero-Shot In-Context Learning

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 3: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning without chain-of-thought prompting.

Classification Prompt: Zero-Shot In-Context Learning with Implicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response. Let's think step by step.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 4: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning with implicit chain-of-thought prompting.

Classification Prompt: Zero-Shot In-Context Learning with Explicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Perform the below reasoning steps in the same order to correctly respond.

1. Read the provided caption of a scientific figure and locate the possible options for the answer, including both explicitly stated options and options implied in the caption.
2. Look at the provided image of a scientific figures and locate the possible options for the answer, including both explicitly stated options and options implied in the image.
3. Aggregate your located answer options from the caption and image of a scientific figure.
4. Select unique options for the answer from your aggregated options and remove the 'h) none' option if other unique options are present.
5. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 5: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning with explicit chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

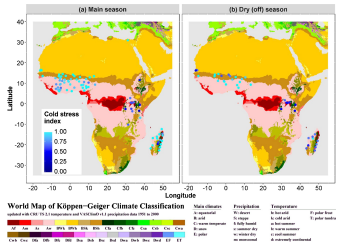


Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The response is: [‘precipitation’, ‘temperature’]

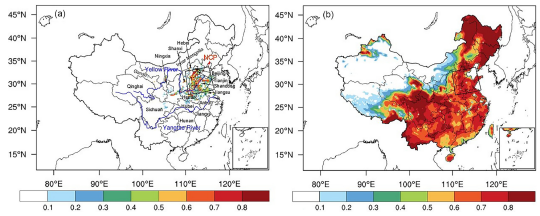


Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The response is: [‘none’]

RESPONSE:

Figure 6: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning without chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning with Implicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response. Let’s think step by step.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

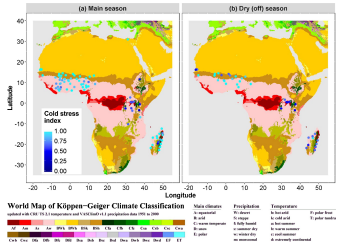


Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The response is: [‘precipitation’, ‘temperature’]

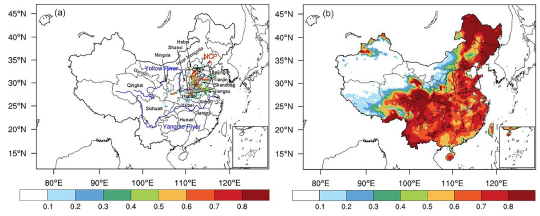


Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The response is: [‘none’]

RESPONSE:

Figure 7: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning with implicit chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning with Explicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Perform the below reasoning steps in the same order to correctly respond.

1. Read the provided caption of a scientific figure and locate the possible options for the answer, including both explicitly stated options and options implied in the caption.
2. Look at the provided image of a scientific figures and locate the possible options for the answer, including both explicitly stated options and options implied in the image.
3. Aggregate your located answer options from the caption and image of a scientific figure.
4. Select unique options for the answer from your aggregated options and remove the 'h) none' option if other unique options are present.
5. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

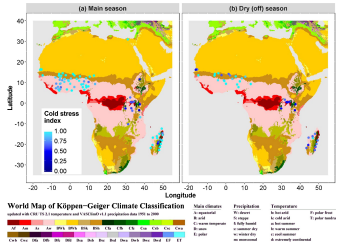


Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The performed reasoning steps are:

1. The possible options for the answers are 'precipitation' and 'temperature'. Although there are no explicitly stated options when reading the caption, the 'temperature' is implied in the mention of 'cold stress' climate science term that represents strain caused by prolonged exposure to low temperatures and temperature drops. Also, both 'temperature' and 'precipitation' is implied in the 'Köppen–Geiger' climate science term that classifies Earth's climates based on temperature and precipitation.
2. The possible options for the answers are 'precipitation' and 'temperature', which are both explicitly stated options when looking at the image.
3. The aggregated answer options are 'precipitation' and 'temperature' from the caption and 'precipitation' and 'temperature' from the image of a scientific figure.
4. The selected unique options for the answer are 'precipitation' and 'temperature' after removing the duplicated options from the aggregated answer options.
5. The response in Python format for a list of items is ['precipitation', 'temperature'].

The response is: ['precipitation', 'temperature']

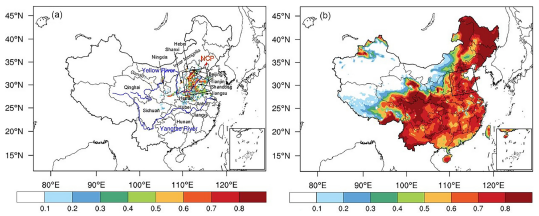


Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The performed reasoning steps are:

1. The possible options for the answers are 'none' as there are no explicitly or implicitly stated options when reading the caption.
2. The possible options for the answers are 'none' as there are no explicitly or implicitly stated options when looking at the image.
3. The aggregated answer options are 'none' from the caption and 'none' from the image of a scientific figure.
4. The selected unique options for the answer are 'none' after removing the duplicated options from the aggregated answer options.
5. The response in Python format for a list of items is ['none'].

The response is: ['none']

RESPONSE:

Figure 8: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning with explicit chain-of-thought prompting.

AquaCrisis: A Multilingual Benchmark for Classifying Water-Utility Public Communication

Vishal Kumar
Faculty of Information Technology and Electrical Engineering, CMVS, University of Oulu
Pentti Kaiteran katu 1
90570 Oulu, Finland
vishal.kumar@oulu.fi

Mourad Oussalah
Faculty of Information Technology and Electrical Engineering, CMVS, University of Oulu
Pentti Kaiteran katu 1
90570 Oulu, Finland
mourad.oussalah@oulu.fi

Abstract

Water utilities communicate both routine service information and time-sensitive public-risk guidance, yet multilingual NLP benchmarks for institution-authored infrastructure communication remain limited. We introduce AquaCrisis, a corpus of 20,000 text-bearing Facebook posts from 12 European water utilities, with a 1,500-post manually annotated gold standard. We formulate Task A as classification of a post's main water-utility topic and Task B as classification of its primary communicative function, together with a coarser grouped Task B.1. We compare rule-based and supervised classifiers with four instruction-tuned LLMs on native-language and English-translated inputs. Supervised approaches achieve the highest point estimates on the fine-grained tasks: translated TF-IDF with logistic regression performs best on Task A, while translated multilingual sentence embeddings perform best on Task B. GPT-4o-mini achieves the highest point estimate on grouped Task B.1. Paired bootstrap analysis shows that translation effects are model- and setting-dependent: gains are clearest for the lexical TF-IDF classifier and selected Gemma 4 and GPT-4.1-mini conditions, while multilingual sentence embeddings, GPT-4o-mini, and Qwen 3.5 are comparatively stable across input representations. AquaCrisis provides a benchmark for multilingual water-utility communication spanning routine, educational, operational, and incident-related messaging, supporting communication monitoring, operational triage, and cross-lingual evaluation.

1 Introduction

Water utilities are frontline actors in climate adaptation because they connect climate-driven hydrological change with public protective action (Boholm and Prutzer, 2017). Climate change is intensifying water scarcity, drought, rainfall extremes, flooding, salinity, contamination risk, and pressure on urban water infrastructure (Boholm and Prutzer, 2017; Bradford et al., 2017). At the same time, utilities increasingly use social media, including Facebook, to communicate routine updates, water-quality information, public engagement, and emergency messages (Sallaku et al., 2017; Martin, 2017; Mix et al., 2020). Importantly, incident-related messages occur within a much larger stream of routine, educational, customer-service, and community-facing communication.

This communication matters because failures in water-risk messaging can have direct public-health and adaptation consequences (Day et al., 2024; Bradford et al., 2017; Rundblad et al., 2010). Studies of "Do Not Drink" and "Boil Water" notices show that unclear messages can lead to confusion and non-compliance, while effective communication helps the public take protective action during contamination events, outages, and compound disasters (Day et al., 2024; Rundblad et al., 2010). In a warming climate, timely and understandable communication about contamination, disruption, conservation, and water stress is therefore part of adaptation capacity, not only public relations (Boholm and Prutzer, 2017; Mix et al., 2020; Day et al., 2024; Bradford et al., 2017).

Water-utility communication both describes an issue affecting the water system and performs a communicative action toward the public. We therefore formulate the problem as two complementary tasks: Task A identifies what a post is about, such as a water-utility topic, supporting incident monitoring, operational routing, trend analysis, and separation of urgent incidents from routine communication; Task B identifies what the utility is trying to do, such as warn, instruct, update, or reassure the public, supporting communication audits and tracking how incidents are communicated over time. The tasks are related but not interchangeable: a burst-pipe post may concern service disruption under Task A, while its Task B function may be a warning, an instruction to store water, a repair

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 246–281
October 28, 2026 ©2026 Association for Computational Linguistics

update, or reassurance that service has been restored. Task A therefore captures the subject of the message, whereas Task B captures its communicative purpose, together supporting both operational monitoring and evaluation of public-risk communication.

Natural language processing (NLP) offers a way to analyze such communication at scale, but existing work does not yet fully address this setting (Bonenkamp et al., 2025; Kurfali et al., 2025). Crisis-NLP has developed datasets and methods for relevance, informativeness, actionability, and urgency in disaster social media, including multilingual settings (Kaufhold et al., 2020; Al Amer et al., 2023; Meunier et al., 2025). ClimateNLP has expanded benchmarks for climate-related classification, retrieval, and intercultural communication (Bonenkamp et al., 2025; Kurfali et al., 2025; Bird et al., 2024). However, institution-authored infrastructure communication differs from citizen-generated crisis streams: most messages are routine or educational, while warnings and operational incidents form a smaller but important subset. To our knowledge, existing multilingual benchmarks do not jointly model the main topic and communicative function of European water-utility social-media communication.

To address this gap, we introduce a multilingual dataset of Facebook posts from European water utilities and evaluate NLP models and large language models (LLMs) to classify both the main water-utility topic and the primary communicative function of each post. We ask:

RQ1: How well can NLP models and LLMs classify multilingual water-utility Facebook posts by main topic and primary communicative function?

RQ2: How does classification performance differ between native-language and English-translated inputs, and how does the effect of translation vary across languages?

RQ3: How do LLMs compare with classical baselines such as keyword rules, term frequency–inverse document frequency (TF-IDF), and multilingual embedding models?

Our label space covers water quality, contamination, service disruption, conservation, drought, flooding, wastewater, and related public communication, capturing both acute risk messaging and longer-term climate-water-stress communication (Lyle et al., 2023; Nijhawan and Howard, 2022; Schewe et al., 2014). We evaluate models on native-language posts and English-translated inputs, motivated by evidence that translation-based multilingual evaluation may help some models but can also distort source-language performance (Al Amer et al., 2023; Enomoto et al., 2025; Hu et al., 2020).

Our contributions are: (1) a multilingual corpus of 20,000 institution-authored water-utility posts with a 1,500-post manually annotated benchmark; (2) complementary topic and communication-function classification tasks; and (3) a systematic comparison of rule-based, supervised, and prompted LLM approaches across native and translated inputs.

2 Related Studies

ClimateNLP has developed a growing set of domain-specific datasets and benchmarks for climate claims, adaptation and policy documents, sustainability disclosures, and scientific information extraction (Bonenkamp et al., 2025; Kurfali et al., 2025; Diggelmann et al., 2020; Morio and Manning, 2023; Pan et al., 2025; Stede and Patz, 2021; Planas et al., 2022; Ni et al., 2023; Schimanski et al., 2024). These resources demonstrate the value of domain-specific classification and information extraction, but they predominantly target reports, scientific documents, policy texts, or corporate claims. AquaCrisis instead focuses on short, institution-authored messages from infrastructure providers, where climate-sensitive topics such as drought and flooding coexist with operational and routine public communication.

Crisis informatics shows that social media can support situational awareness, emergency response, and public communication (Kaufhold et al., 2020; Alam et al., 2018; Kumar et al., 2025; Alam et al., 2021a,b). Related work has also applied text-mining methods to mobile emergency-communication applications, showing how user feedback can reveal communication needs and support early-warning and emergency-management systems (Shaik and Oussalah, 2024). CrisisBench, HumAID, CrisisMMD, cross-lingual crisis classification, and SPEED++ address informativeness, humanitarian categories, multimodal understanding, and multilingual event extraction (Al Amer et al., 2023; Alam et al., 2018, 2021a,b; Parekh et al., 2024). Instruction-tuned LLMs can also

perform competitively on crisis-post classification, although supervised baselines often remain stronger (McDaniel et al., 2024). Water-related social-media NLP is more limited: WaterMM introduced bilingual water-quality relevance classification, while later studies examined water-quality feedback, contamination discourse, public opinion, and surveillance (Andreadis et al., 2022; Ahmad et al., 2023; Ayub et al., 2022; Mix et al., 2020). Prior work has mainly focused on citizen-generated Twitter/X content or binary relevance filtering, leaving institution-authored Facebook communication and fine-grained communicative functions understudied (Sallaku et al., 2017; Martin, 2017; Cheng et al., 2022).

Our taxonomy also draws on research on drinking-water risk and crisis communication. Work on drinking-water notices distinguishes the underlying hazard or operational issue from the communicative action directed toward the public, such as warning, instructing, explaining, or reassuring (Day et al., 2024; Bradford et al., 2017; Rundblad et al., 2010). Social-mediated crisis-communication research likewise emphasizes that the form and purpose of institutional messages shape public response (Jin et al., 2014; Cheng et al., 2022). AquaCrisis does not directly implement a single existing theoretical taxonomy; rather, its labels synthesize these communication perspectives with water-utility practice and pilot inspection of the corpus. Task A captures the dominant subject of a post, while Task B captures its primary communicative action, allowing the same operational topic to be studied across warning, instruction, update, reassurance, education, and institutional communication.

Our work also builds on multilingual NLP and translation-based inference (Enomoto et al., 2025; Hu et al., 2020; Conneau et al., 2020; Barbieri et al., 2022; Liu et al., 2025). Multilingual encoders such as XLM-R and XLM-T provide strong representations for noisy cross-lingual short-text classification (Conneau et al., 2020; Barbieri et al., 2022). Translating inputs into English can benefit English-oriented models, but may lose linguistic or cultural nuance, and native-language specialization can outperform translation-based approaches (Enomoto et al., 2025; Liu et al., 2025; Mullappilly et al., 2023). These findings motivate our comparison of native-language and English-translated water-utility posts rather than assuming a single multilingual strategy is optimal.

3 Methodology

Figure 1 summarizes the AquaCrisis corpus-construction and evaluation workflow. We first collect and preprocess multilingual posts from public Facebook pages operated by water utilities. Each post is represented in two forms: its native-language text and an English translation produced with a zero-shot GPT-4.1-nano prompt. Details of the preprocessing and translation process are provided in Appendix A. A stratified sample of 1,500 posts is then selected for gold-label annotation under Task A (Main water-utility topic) and Task B (Primary communicative function). The resulting gold-standard dataset is used to evaluate both supervised baselines and LLMs. The experiments compare native and translated text representations and include rule-based methods, stratified five-fold cross-validation for supervised models, and zero- and five-shot prompting for LLMs.

3.1 Data collection and corpus construction

Table 1 summarizes the original Facebook collection. We construct a multilingual corpus of public Facebook posts from water utilities and water-service organizations in European cities, collected with Apify (Apify, 2026). The original collection contains 29,352 Facebook records from 12 public utility pages, spanning August 2009 to June 2026 and covering 11 language labels. After deduplicating by post ID and removing records without usable textual content, the final corpus contains 20,000 posts. For each post, we retain the original text, timestamp, source URL, media type, and engagement metadata, including reactions, comments, and shares. We focus on Facebook because utilities use it for longer-form public updates, operational notices, and community-facing communication. Details of the utility accounts are provided in Table 5 of Appendix D.

We model these posts as public communication about water quality, service disruption, drought and conservation, flooding and wastewater, and related utility functions, rather than as generic organizational text. This follows crisis-informatics work that labels social-media posts for operational categories such as informativeness, humanitarian need, and event structure (Alam et al., 2018, 2021a,b; Imran et al., 2016). It also places the dataset within ClimateNLP, where recent work emphasizes domain-specific corpora, climate benchmarks, and adaptation-oriented tasks (Kurfali et al., 2025; Liu

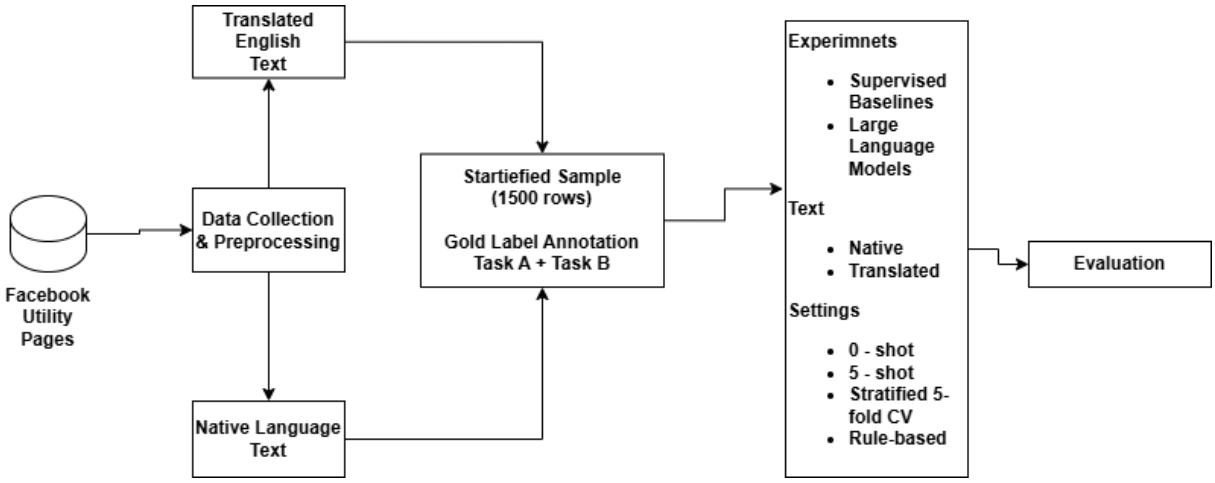


Figure 1: Overview of the AquaCrisis corpus-construction and evaluation workflow

Statistic	Original Facebook collection
Raw collected rows	29,352
Unique post IDs	20,883
Utility pages	12
Platform	Facebook
Date range	6 Aug. 2009–8 Jun. 2026
Explicit language labels	11 languages
Final text-bearing posts	20,000

Table 1: Statistics of the original collected Facebook dataset

et al., 2025; Webersinke et al., 2021). Our corpus extends this agenda to climate-sensitive infrastructure communication, where utilities communicate risk, reassurance, instruction, and service continuity to the public. The corpus is unevenly distributed across languages and utilities; Portuguese and Finnish together account for nearly half of the preprocessed posts, while several languages contribute substantially smaller subsets (Figure 3).

3.2 Annotation scheme and gold standard

We annotate a 1,500-post gold-standard subset in Label Studio (HumanSignal, 2026). Instead of uniform random sampling, we use a stratified, keyword-guided strategy to increase coverage of rare but operationally important posts. The keyword lists were derived from the same domain-informed taxonomy used for the classification tasks: we first defined categories from water-utility communication and climate-related water-risk knowledge, then selected keywords reflecting those categories. This follows crisis-dataset practice, where task-aware sampling helps capture informative but infrequent crisis content (Alam et al., 2021a). The full sampling lexicon and priority signals are reported in Table 8 in Appendix G. The resulting gold set is therefore an enriched benchmark sample rather than an estimate of category prevalence in the full corpus. Because sampling uses taxonomy-related lexical cues on English-translated text, performance, particularly for lexical methods such as translated TF-IDF, may also be affected by this enrichment.

Two domain-informed annotators reviewed the guidelines and completed a pilot round to clarify label definitions and difficult cases. They then independently assigned one Task A and one Task B label to each post. Before consensus, raw inter-annotator agreement was 72.3% for Task A and 79.1% for Task B, with Cohen's κ of 0.672 and 0.715, respectively. Disagreements were resolved through discussion and consensus; unresolved cases were adjudicated by a third domain-informed annotator using the post, proposed labels, and guidelines. Annotators read the English translation and consulted the original-language text when the translation was unclear or potentially misleading. Detailed agreement statistics are reported in Table 4 of Appendix C.1, with the full annotation guidelines provided in Appendix C.

We define two main tasks. **Task A** asks, "What is the post mainly about?" and assigns the post's primary water-utility topic: Water quality / safety / public health; Service disruption / repair / infrastructure; Water conservation / drought / demand management; Flood / stormwater / wastewater /

sewer; Environmental sustainability / waste / recycling / biodiversity; Public education / heritage / community engagement; and Routine / institutional / customer service / other. **Task B** asks, "What is the utility mainly trying to do through the post?" and assigns its primary communicative function: Alert / warning; Instruction / advice to public; Operational update / resolution; Reassurance / safety information; Education / awareness; Institutional promotion / community news; and Other / unclear.

Real posts may contain several topics or communicative functions. We nevertheless use one primary label per task to create a well-defined retrieval and routing benchmark: annotators select the topic that motivates the post or represents its main consequence, and the communicative function receiving the greatest emphasis or carrying the most important public-risk implication. This simplifies multi-function messages and should therefore be viewed as a primary-label representation rather than an exhaustive discourse annotation; multi-label annotation is a natural extension.

Because alerts, instructions, operational updates, and reassurance frequently overlap conceptually, we additionally derive **Task B.1** by mapping Alert / warning, Instruction / advice to public, Operational update / resolution, and Reassurance / safety information to Risk / incident communication while retaining education, institutional/community communication, and other/unclear as separate groups. B.1 is derived deterministically from Task B and provides a coarser evaluation of communicative role.

3.3 Baselines and multilingual classifiers

We evaluate the supervised baselines using stratified five-fold cross-validation on the annotated gold standard. In each fold, models are trained on approximately 80% of the labeled posts and evaluated on the held-out 20%, while preserving the label distribution where possible. We pool the out-of-fold predictions so that every annotated post is evaluated exactly once by a model that was not trained on that post. This provides a complete gold-set evaluation for comparison with the LLM results and reduces dependence on a single train-test split.

We include a majority-class baseline as a lower bound and a keyword-rule baseline as an interpretable monitoring approach. The keyword system assigns labels using manually curated lexical triggers associated with the task taxonomy and falls back to a default routine/other label when no rule matches. The complete Task A and Task B lexicons are reported in Tables 9 and 10 in Appendix G.

Our second baseline is TF-IDF + logistic regression (LR). We compute unigram and bigram TF-IDF features and train a class-balanced logistic regression classifier on native and translated text variants. This setting is useful because short utility posts often contain recurring phrases such as water interruption, boil water, planned maintenance, heavy rain, or save water. Classical text classifiers remain strong references for short-text classification because they are fast, interpretable, and competitive on many domain-specific tasks (Joulin et al., 2017).

Our third baseline is a multilingual sentence-embedding classifier. We encode posts with paraphrase-multilingual-MiniLM-L12-v2 from Sentence-Transformers and train logistic regression on the dense embeddings. This follows the Sentence-BERT line of work, where semantically aligned sentence representations improve generalization across paraphrases and languages compared with purely lexical features (Reimers and Gurevych, 2019). Detailed baseline parameters are reported in Table 6 of Appendix E.

3.4 Zero-shot and few-shot LLM classification

We evaluate instruction-following LLMs with zero-shot and five-shot prompting on both original-language and English-translated posts. We use Gemma 4 (Gemma Team, 2026), GPT-4.1-mini (OpenAI, 2025), GPT-4o-mini (OpenAI, 2024), and Qwen 3.5 (Qwen Team, 2026) because they support multilingual inputs and enable direct prompted classification. The GPT-4o-mini setting is motivated by recent crisis-social-media work evaluating GPT-4o-family models for zero-shot crisis-tweet classification on CrisisBench, but we use the smaller model as a cost-efficient alternative for our larger native/translated and zero/five-shot benchmark design (McDaniel et al., 2024). For each post, the model predicts one Task A label and one Task B label. The prompt also requests a self-reported confidence value; this was collected as an exploratory diagnostic and was not used for prediction selection, filtering, or evaluation. We requested structured JSON output, but did not use grammar-constrained or guided decoding beyond the backend's JSON-formatting mechanism.

This setup connects ClimateNLP and crisis-social-media NLP: the task is domain-specific, multilingual, and operational, but the adaptation

method is prompting rather than task-specific fine-tuning (Bonenkamp et al., 2025; Kurfali et al., 2025; Pan et al., 2025; Alam et al., 2021b). Zero-shot and few-shot prompting are standard adaptation settings for modern language models, but prior work shows that demonstrations, prompt order, and calibration can substantially affect stability (Zhao et al., 2021). We therefore treat five-shot prompting as a controlled experimental condition rather than assuming that adding examples will always improve performance. The full zero-shot and five-shot prompts are shown in Appendix B along with the prompt design details. Exact model identifiers, inference backends, and decoding parameters are reported in Table 7 of Appendix F.

3.5 Evaluation protocol

We report accuracy, macro-F1, weighted-F1, and per-class precision, recall, and F1 for both tasks. For the supervised baselines, the primary reported scores are computed from the pooled out-of-fold predictions across the five folds. Macro-F1 is central because the corpus contains frequent routine posts alongside rare but important categories such as contamination, service disruption, and warnings. Weighted-F1 complements this by reflecting performance under the observed label distribution. This metric bundle follows recent crisis-social-media benchmark practice (Al Amer et al., 2023; Alam et al., 2021a,b; Parekh et al., 2024). For supervised baselines, fold-level macro-F1 means and standard deviations are also provided as a complementary stability analysis.

Invalid generative predictions are retained rather than discarded. An invalid or out-of-schema Task A prediction is scored as incorrect for Task A, while an invalid Task B prediction is scored as incorrect for Task B and the derived Task B.1. When only one task prediction is invalid, the valid prediction for the other task is still evaluated normally. A malformed confidence value alone does not invalidate otherwise valid task labels. No gold examples are removed because of invalid model output. We additionally report the proportion of rows containing at least one invalid task label as a schema-compliance diagnostic.

To quantify uncertainty, we compute 95% percentile bootstrap confidence intervals using 10,000 stratified-bootstrap replicates. For each task, posts are resampled with replacement within the corresponding gold-label strata, and identical resampled post indices are used across systems for paired comparisons. Macro-F1 is recomputed over the predefined label set on every replicate, with invalid predictions retained as incorrect. We describe evidence for a difference when the 95% paired interval for Δ macro-F1 excludes zero. Fold-level means and standard deviations for the supervised baselines are retained as a separate measure of cross-validation stability.

4 Results

4.1 Label distribution

Table 2 shows substantial class imbalance. Routine and institutional communication is the largest Task A category, while institutional news and education dominate Task B. Operational updates, reassurance, and unclear posts are especially rare, motivating our use of macro-F1 and the grouped Task B.1 evaluation in Table 3. The skew also helps explain why Task B is generally more difficult than Task A and why grouped Task B.1 improves performance.

4.2 Model performance

Table 3 reports macro-F1 scores on the gold standard. The majority baseline performs poorly (Task A: 0.069; Task B: 0.085), reflecting class imbalance. Keyword rules provide a stronger but limited baseline (0.562/0.309), while supervised models perform best on the fine-grained tasks. TF-IDF + LR on translated text achieves the highest Task A score (0.671), while multilingual embeddings + LR on translated text achieve the highest Task B score (0.504). Among LLMs, Gemma 4 reaches 0.658 on Task A, GPT-4.1-mini reaches 0.638, and GPT-4o-mini reaches 0.624. For Task B, GPT-4o-mini performs best among LLMs, reaching 0.392 with translated zero-shot prompting. Overall, task-specific supervised models remain strongest for both tasks A and B. Bootstrap intervals for individual systems are reported in Appendix H.4; paired comparisons are used below when interpreting differences between input conditions.

Task B is generally more difficult than Task A across the evaluated systems, plausibly because its labels are more ambiguous and imbalanced. Grouping sparse and overlapping Task B classes into Task B.1 improves most scores and particularly benefits GPT-4o-mini. Translated multilingual embeddings + LR rise from 0.504 to 0.592, while GPT-4o-mini reaches 0.660 with translated five-shot prompting and 0.650 with native five-shot prompting. This indicates that broader communicative-

Task	Category	Count	%
Task A	Water quality / safety / public health	65	4.3
Task A	Service disruption / repair / infrastructure	226	15.1
Task A	Water conservation / drought / demand management	140	9.3
Task A	Flood / stormwater / wastewater / sewer	136	9.1
Task A	Environmental sustainability / waste / recycling / biodiversity	209	13.9
Task A	Public education / heritage / community engagement	250	16.7
Task A	Routine / institutional / customer service / other	474	31.6
Task B	Alert / warning	205	13.7
Task B	Instruction / advice to public	164	10.9
Task B	Operational update / resolution	21	1.4
Task B	Reassurance / safety information	11	0.7
Task B	Education / awareness	446	29.7
Task B	Institutional promotion / community news	638	42.5
Task B	Other / unclear	15	1.0

Table 2: Label distribution in the 1,500-post gold-standard annotation sample for Task A and Task B.

Model	Text	Setting	Task A	Task B	Task B.1	Invalid Rows %
Majority	all	Stratified 5-fold CV	0.069	0.085	0.150	–
Keyword rules	translated	rule-based	0.562	0.309	0.356	–
TF-IDF + LR	translated	Stratified 5-fold CV	**0.671**	0.475	0.548	–
TF-IDF + LR	native	Stratified 5-fold CV	0.521	0.394	0.475	–
Multilingual embeddings + LR	translated	Stratified 5-fold CV	0.657	**0.504**	0.592	–
Multilingual embeddings + LR	native	Stratified 5-fold CV	0.649	0.495	0.575	–
Gemma 4	translated	zero-shot	0.658	0.352	0.563	0.00
Gemma 4	translated	five-shot	0.624	0.331	0.553	12.27
Gemma 4	native	zero-shot	0.633	0.346	0.552	0.00
Gemma 4	native	five-shot	0.566	0.292	0.497	19.80
GPT-4.1-mini	translated	zero-shot	0.638	0.359	0.494	0.00
GPT-4.1-mini	translated	five-shot	0.617	0.325	0.466	0.00
GPT-4.1-mini	native	zero-shot	0.616	0.328	0.491	0.07
GPT-4.1-mini	native	five-shot	0.596	0.315	0.463	0.00
GPT-4o-mini	translated	zero-shot	0.620	0.392	0.646	0.00
GPT-4o-mini	translated	five-shot	0.624	0.381	**0.660**	0.00
GPT-4o-mini	native	zero-shot	0.601	0.384	0.623	0.00
GPT-4o-mini	native	five-shot	0.605	0.372	0.650	0.00
Qwen 3.5	translated	zero-shot	0.567	0.277	0.472	0.13
Qwen 3.5	translated	five-shot	0.584	0.269	0.441	0.13
Qwen 3.5	native	zero-shot	0.558	0.277	0.461	0.53
Qwen 3.5	native	five-shot	0.573	0.242	0.430	1.20

Table 3: Model comparison on the 1,500-post gold standard. Scores are macro-F1. Supervised models are evaluated using pooled out-of-fold predictions from stratified five-fold cross-validation (CV). Keyword rules and LLMs are evaluated on the complete annotated sample. Invalid rows (%) denotes the percentage of LLM outputs with a missing, malformed, or out-of-schema Task A or Task B label. Bootstrap confidence intervals for the LLM systems are reported in Appendix H.4; paired baseline translation effects are reported in Section 4.3.

function categories are more robust than the original fine-grained labels and that LLMs become especially competitive when the label space is less granular. Additional accuracy, weighted-F1, class-wise results, fold-level variability, per-language comparisons, and bootstrap uncertainty analyses are reported in Tables 11–24 and Figures 4–9 of Appendix H.

LLMs also sometimes produced labels outside the predefined schema. The rightmost column of Table 3 reports invalid-output rates. Gemma 4 five-shot is the least stable, with 12.27% invalid outputs on translated text and 19.80% on native text. Qwen 3.5 remains at or below 1.20%, GPT-4.1-mini produces only one invalid native zero-shot row, and GPT-4o-mini produces no invalid rows. Five-shot prompting therefore does not consistently improve either performance or schema compliance. These errors show that generative classification can return missing, malformed, or out-of-schema labels, so macro-F1 should be interpreted together with invalid-output rate, especially for operational use.

4.3 Translation versus native text

Figure 2 visualizes the native-versus-translated macro-F1 comparisons and their 95% bootstrap confidence intervals for Tasks A and B. Translation effects are strongly model-dependent. For TF-IDF + LR, translated input increases macro-F1 by 0.150 on Task A (95% paired confidence interval (CI) [0.124, 0.175]), by 0.080 on Task B ([0.037, 0.127]), and by 0.073 on Task B.1 ([0.055, 0.091]). In contrast, the translated–native differences for multilingual embeddings are small and their paired intervals include zero on all three tasks.

Among LLMs, translated input is higher for

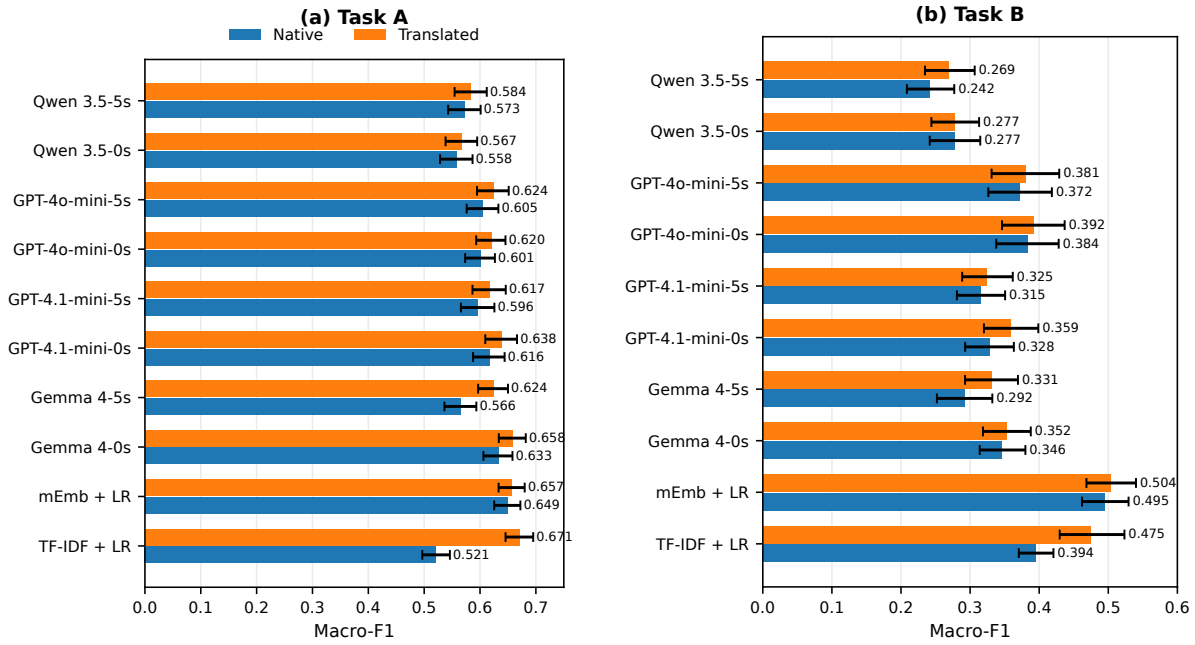


Figure 2: Native versus translated model performance on Tasks A and B. Bars show macro-F1 and whiskers show 95% percentile bootstrap confidence intervals from 10,000 stratified-bootstrap replicates. 0s and 5s denote zero-shot and five-shot prompting; mEmb = multilingual embeddings; LR = logistic regression. Paired LLM translated–native confidence intervals are reported in Appendix H.4

Gemma 4 on Task A in zero-shot prompting ($\Delta = 0.025$, [0.007, 0.045]) and on all three tasks under five-shot prompting ($\Delta = 0.058$, [0.033, 0.085] for Task A; 0.039, [0.004, 0.076] for Task B; and 0.056, [0.019, 0.092] for Task B.1). GPT-4.1-mini zero-shot also shows translated-input differences on Task A (0.022, [0.002, 0.043]) and Task B (0.032, [0.010, 0.054]). All translated–native intervals for GPT-4o-mini and Qwen 3.5 include zero. Translation therefore provides clear gains for the lexical classifier and selected LLM conditions, rather than a uniform advantage across models.

Per-language results are also heterogeneous (Appendix H.3). Among the eight languages with at least 50 gold posts, translated input gives numerically higher zero-shot Task A macro-F1 for Gemma 4 in all eight languages and for GPT-4.1-mini in seven of eight, whereas directions are less consistent for GPT-4o-mini and Qwen 3.5. Because several language subsets remain small and class coverage differs, we treat these results descriptively rather than as language-level significance tests.

5 Discussion

For RQ1, multilingual water-utility posts are classifiable, but performance depends strongly on the task. Task A yields higher scores than Task B for all non-majority systems in our experiments, which may reflect the more concrete nature of the topic labels relative to the communicative-function distinctions. Translated TF-IDF has the highest Task A score, while translated multilingual embeddings perform best for Task B, showing that both lexical normalization and cross-lingual semantic representations are effective for short, domain-specific posts (Conneau et al., 2020; Barbieri et al., 2022; Stammbach et al., 2023; Schimanski et al., 2023). The improvement on grouped Task B.1 is consistent with broader communicative roles being easier to classify reliably than the fine-grained distinctions, particularly for rare labels such as reassurance and operational resolution (Mix et al., 2020; Andreadis et al., 2022; Stede and Patz, 2021; Schimanski et al., 2024).

For RQ2, translation effects are model- and setting-dependent rather than uniformly positive. English translation substantially benefits TF-IDF, consistent with lexical normalization across languages, whereas the multilingual embedding classifier shows small paired differences whose intervals include zero. Among LLMs, translated input shows evidence of gains for selected Gemma 4 and GPT-4.1-mini conditions, while the corresponding intervals for GPT-4o-mini and Qwen 3.5 in-

clude zero. The language-specific analysis further shows heterogeneous patterns across the eight sufficiently represented languages. These results support using translation as an experimental alternative rather than replacing native-language evaluation (Enomoto et al., 2025; Liu et al., 2025; Mullappilly et al., 2023).

For RQ3, LLMs are competitive but do not consistently outperform supervised baselines. Supervised models remain strongest on the fine-grained tasks, whereas GPT-4o-mini has the highest Task B.1 score, reaching 0.660 with translated five-shot prompting compared with 0.592 for the strongest supervised baseline. This suggests that instruction-tuned LLMs are particularly effective when the label space is broader, while task-specific classifiers retain an advantage for finer distinctions (McDaniel et al., 2024; Kurfali et al., 2025; Ni et al., 2023).

The LLM results also reveal practical limitations. Five-shot prompting does not provide consistent gains. For Gemma 4, five-shot Task A is lower than zero-shot on both native ($\Delta = -0.067$, 95% CI [-0.094, -0.041]) and translated input (-0.034, [-0.055, -0.013]); several GPT-4.1-mini and Qwen comparisons are likewise lower, whereas all GPT-4o-mini five-shot–zero-shot intervals include zero. Nevertheless, generative classification remains sensitive to formatting and can return missing, malformed, or out-of-schema labels unless outputs are tightly constrained (McDaniel et al., 2024; Enomoto et al., 2025; Ni et al., 2023).

Water-utility communication remains challenging because posts are short, local, technical, multilingual, and often combine several communicative purposes. A single message may include infrastructure information, service status, advice, and reassurance, while small wording differences can carry important public-health consequences (Bradford et al., 2017; Rundblad et al., 2010). The main practical value of automated classification is therefore in triage and monitoring rather than full automation. Models can help identify alerts, contamination risks, drought, flooding, and service disruption, supporting situational awareness (Alam et al., 2018, 2021a,b; Planas et al., 2022).

Human oversight remains essential. Misclassifying a warning as routine, or reassurance as unrelated content, could affect public understanding, compliance, and trust (Day et al., 2024; Bradford et al., 2017; Rundblad et al., 2010). NLP should therefore be treated as decision support: models can filter and prioritize posts, but expert communicators should verify high-stakes cases. Overall, this study extends climate and crisis NLP to multilingual, institution-authored water-utility communication while keeping human judgment central (Andreadis et al., 2022; Ahmad et al., 2023; Ayub et al., 2022; Jin et al., 2014; Stede and Patz, 2021; Planas et al., 2022; Schimanski et al., 2024; Kumar et al., 2025).

6 Conclusion

We introduced a multilingual benchmark for classifying Facebook communication from European water utilities by topic and communicative function. The dataset captures both routine institutional messaging and climate-relevant communication about water quality, service disruption, drought, flooding, wastewater, sustainability, and public guidance. We evaluate fine-grained Task A and Task B labels together with the broader grouped Task B.1 categories.

The results show that water-utility topics are easier to classify than fine-grained communicative functions. Translated TF-IDF with logistic regression performs best on Task A, while translated multilingual sentence embeddings with logistic regression perform best on Task B. LLMs remain competitive but do not consistently surpass supervised baselines; GPT-4o-mini performs best on grouped Task B.1, reaching 0.660 with translated five-shot prompting. Translation strongly benefits the lexical TF-IDF classifier, whereas paired uncertainty estimates show more selective effects for LLMs: some Gemma 4 and GPT-4.1-mini conditions improve with translation, while GPT-4o-mini and Qwen 3.5 are comparatively stable. Generative classification also requires careful schema control, as Gemma 4 produces substantial invalid-output rates in its five-shot conditions.

Future work should move from post-level classification toward incident-centered analysis by linking related messages into event timelines and extracting affected locations, incident status, public advice, and resolution information. Connecting utility communication with water-quality, weather, sensor, and emergency records could support monitoring while retaining expert oversight for high-stakes interpretation.

Limitations

This study has several limitations. The corpus is limited to Facebook posts from European wa-

ter utilities. Facebook is useful for studying public-facing updates and community engagement, but it does not capture the full utility communication ecosystem, which may include websites, press releases, emergency-alert systems, X/Twitter, LinkedIn, SMS alerts, and local media. The results therefore reflect Facebook-based utility communication in European contexts and may not generalize to regions with different infrastructure, regulation, climate risks, or communication practices.

The corpus is not balanced across languages or utilities. Some languages and pages contribute many more posts than others, reflecting real activity patterns rather than a balanced multilingual benchmark design. This may affect supervised learning and evaluation, especially for lower-frequency languages, and means that native-versus-translated results should not be interpreted as equally strong evidence for every language.

Emergency and outbreak-related posts are rare in the corpus. Most posts concern routine updates, education, community engagement, or customer service, while contamination events, boil-water notices, urgent warnings, reassurance messages, and operational resolutions occur less often. These rare categories are highly important for crisis communication, but their low frequency limits the reliability of per-class estimates.

The translated condition depends on GPT-4.1-nano machine translation. Translation enables comparison between native-language and English-translated inputs, but it may introduce lexical, cultural, or pragmatic errors, especially for local infrastructure terms, place names, informal language, and risk-communication phrasing. Improvements on translated text should therefore not be read as evidence that translation is always preferable. Because annotators primarily reviewed the English translations and consulted the original-language text when needed, translation errors may also have influenced some gold labels, introducing potential annotation bias.

The manually annotated gold standard contains 1,500 posts. This is sufficient for an initial benchmark and model comparison, but it limits reliability for rare labels, and macro-F1 remains sensitive to minority-class performance. Future work should expand the annotation set, include more incident-focused examples, and consider hierarchical or multi-label schemes for posts combining risk topics, instructions, reassurance, and institutional communication.

Language-specific estimates are uneven in reliability because the gold set is dominated by a subset of languages and several languages contain few examples of rare Task B classes.

Finally, generative models were evaluated using JSON-formatting constraints but not stronger grammar-constrained or guided decoding; such methods may reduce schema violations and should be evaluated in future work.

Ethical Considerations

The corpus is restricted to public organizational posts made by institutional utility accounts, not private user data. We do not profile individual citizens, infer sensitive traits, or release comment-level personal data. Nevertheless, the corpus may contain place names, infrastructure details, and references to local incidents. The distribution of languages, utilities, and communication categories is uneven, and the dataset should not be treated as representative of all water utilities, regions, or communication platforms.

Data Availability The gold annotations, predefined cross-validation fold assignments, experimental code, prompts, model predictions, and derived evaluation results are publicly available at `https://github.com/vishal310197/Aquacrisis`.

References

K. Ahmad, M. A. Ayub, K. Ahmad, J. Khan, Nasir Ahmad, and A. Al-Fuqaha. 2023. Social media as an instant source of feedback on water quality. *IEEE Transactions on Technology and Society*, 4(4):325–333.

Shareefa Al Amer, Mark Lee, and Phillip Smith. 2023. Cross-lingual classification of crisis-related tweets using machine translation. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 22–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 465–473.

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021a. HumAID: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 933–942.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021b. CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 923–932.

Stelios Andreadis, Ilias Gialampoukidis, Aristeidis Bozas, Anastasia Moumtzidou, Roberto Fiorin, Francesca Lombardo, Anastasios Karakostas, Daniele Norbiato, Stefanos Vrochidis, Michele Ferri, and Ioannis Kompatsiaris. 2022. WaterMM: Water quality in social multimedia task at MediaEval 2021. In *Working Notes Proceedings of the MediaEval 2021 Workshop*, volume 3181 of *CEUR Workshop Proceedings*, Online. CEUR-WS.org.

Apify. 2026. Apify: The largest marketplace of trusted tools for ai. `https://apify.com/`.

Muhammad Asif Ayub, Khubaib Ahmad, Kashif Ahmad, Nasir Ahmad, and Ala Al-Fuqaha. 2022. NLP techniques for water quality analysis in social media content. In *Working Notes Proceedings of the MediaEval 2021 Workshop*, volume 3181 of *CEUR Workshop Proceedings*, Online. CEUR-WS.org.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Steven Bird, Angelina Aquino, and Ian Gumbula. 2024. Envisioning NLP for intercultural climate communication. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 111–122, Bangkok, Thailand. Association for Computational Linguistics.

Åsa Boholm and Madeleine Prutzer. 2017. Experts' understandings of drinking water risk management in a climate change scenario. *Climate Risk Management*, 16:133–144.

Jetske Bonenkamp, Robbert Biesbroek, and Ioannis N. Athanasiadis. 2025. Transforming adaptation tracking: Benchmarking transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 266–277, Vienna, Austria. Association for Computational Linguistics.

Lori E. A. Bradford, Blessing Idowu, Rebecca Zagozewski, and Lalita A. Bharadwaj. 2017. There is no publicity like word of mouth... lessons for communicating drinking water risks in the urban setting. *Sustainable Cities and Society*, 29:23–40.

Yang Cheng, Yijing Wang, and Yanqiu Kong. 2022. The state of social-mediated crisis communication research through the lens of global scholars: An updated assessment. *Public Relations Review*, 48(2):102172.

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.

Ashleigh M. Day, Sydney O'Shay, Khairul Islam, Matthew W. Seeger, F. Gianluca Sperone, and Shawn P. McElmurry. 2024. Boil water notices as health-risk communication: Risk perceptions, efficacy, and compliance during Winter Storm Uri. *Scientific Reports*, 14:850.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

HumanSignal. 2026. Label Studio: Open source data labeling platform. `https://labelstud.io/`.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia. European Language Resources Association (ELRA).

Yan Jin, Brooke Fisher Liu, and Lucinda L. Austin. 2014. Examining the role of social media in effective crisis management: The effects of crisis origin, information form, and source on publics' crisis responses. *Communication Research*, 41(1):74–94.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Marc-André Kaufhold, Nicola Rupp, Christian Reuter, and Matthias Habdank. 2020. Mitigating information overload in social media during conflicts and crises: Design and evaluation of a cross-platform alerting system. *Behaviour & Information Technology*, 39(3):319–342.

Vishal Kumar, Frank Hopfgartner, Pekka M. Rossi, and Mourad Oussalah. 2025. Crisis communication and media influence during nokia water contamination event. *Progress in Disaster Science*, 28:100487.

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.

Chia-Hsuan Liu, Yue Zhang, Hao Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Zia J. Lyle, Jeanne M. VanBriesen, and Constantine Samaras. 2023. Drinking water utility-level understanding of climate change effects to system reliability. *ACS ES&T Water*, 3(8):2395–2406.

Sebastian Martin. 2017. When public utilities like to talk on facebook. *Zeitschrift für Energiewirtschaft*, 41(4):311–319.

Emma McDaniel, Samuel Scheele, and Jeff Liu. 2024. Zero-shot classification of crisis tweets using instruction-finetuned large language models. *Preprint*, arXiv:2410.00182.

Romain Meunier, Farah Benamara, Véronique Moriceau, Zhongzheng Qiao, and Savitha Ramasamy. 2025. CrisisTS: Coupling social media textual data and meteorological time series for urgency classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16082–16099, Vienna, Austria. Association for Computational Linguistics.

Nelson Mix, Aaron George, and Adam Haas. 2020. Social media monitoring for water quality surveillance and response systems. *Journal AWWA*, 112(8):44–55.

Gaku Morio and Christopher D. Manning. 2023. An NLP benchmark dataset for assessing corporate climate policy engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.

Sahal Shaji Mullappilly, Sargam Chatterjee, Santosh T.y.s.s, Gokul Narayanan, Hady ElSahar, Bahruz Jabiyev, and Mohamed Ali. 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, Singapore. Association for Computational Linguistics.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

Anisha Nijhawan and Guy Howard. 2022. Associations between climate variables and water quality in low- and middle-income countries: A scoping review. *Water Research*, 210:117996.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI.

OpenAI. 2025. Introducing gpt-4.1 in the api. OpenAI. Introduces GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano.

Huitong Pan, Mustapha Adamu, Qi Zhang, Eduard Dragut, and Longin Jan Latecki. 2025. ClimateIE: A dataset for climate science information extraction. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 76–98, Vienna, Austria. Association for Computational Linguistics.

Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. 2024. SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, USA. Association for Computational Linguistics.

Jordi Planas, Daniel Firebanks-Quevedo, Galina Naydenova, Ramansh Sharma, Cristina Taylor, Kathleen Buckingham, and Rong Fang. 2022. Beyond modeling: NLP pipeline for efficient environmental policy analysis. *arXiv preprint arXiv:2201.07105*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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.

Gabriella Rundblad, Olivia Knapton, and Paul R. Hunter. 2010. Communication, perception and behaviour during a natural disaster involving a Do Not Drink and a subsequent Boil Water notice: A postal questionnaire study. *BMC Public Health*, 10:641.

Rezarta Sallaku, Vania Vigolo, and Federico Testa. 2017. Social media communication in water utilities. In *Proceedings of the 20th Excellence in Services International Conference*, pages 719–730, Verona, Italy.

Jacob Schewe, Jens Heinke, Dieter Gerten, Ingjerd Haddeland, Nigel W. Arnell, Douglas B. Clark, Rutger Dankers, Stephanie Eisner, Balázs M. Fekete, Felipe J. Colón-González, Simon N. Gosling, Hyungjun Kim, Xingcai Liu, Yoshimitsu Masaki, Felix T. Portmann, Yusuke Satoh, Tobias Stacke, Qiuhong Tang, Yoshihide Wada, and 6 others. 2014. Multimodel assessment of water scarcity under climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9):3245–3250.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756, Singapore. Association for Computational Linguistics.

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.

Fuzel Ahamed Shaik and Mourad Oussalah. 2024. On mining mobile emergency communication applications in nordic countries. *International Journal of Disaster Risk Reduction*, 108:104566.

Dominik Stammbach, 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.

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*, pages 8–18, Online. Association for Computational Linguistics.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. ClimateBert: A pretrained language model for climate-related text. *Preprint*, arXiv:2110.12010.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Translation and text representations

Figure 3 shows the language distribution after preprocessing. Each post is represented in two forms: the original-language text and an English translation produced with a zero-shot GPT-4.1-nano prompt. This allows us to compare native-language inference with a translate-to-English pipeline. Because recent multilingual LLM work shows that English translation can help some models but is not uniformly optimal, especially when pragmatic or cultural cues matter, we treat translation as an experimental condition rather than a default preprocessing step (Liu et al., 2025). The full translation prompt can be found in Appendix B.1.

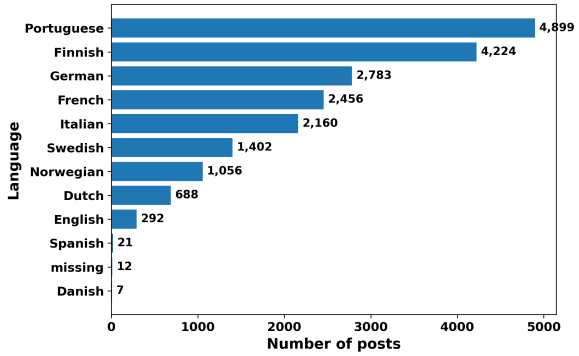


Figure 3: Language distribution in the preprocessed Facebook dataset.

For supervised baselines, we use normalized text to reduce lexical sparsity. The preprocessing pipeline unescapes HTML, normalizes Unicode, removes URLs and mentions, and lowercases text. Hashtag content is preserved as lexical material by converting forms such as #savewater into savewater. For LLM classification, we use the original post strings rather than cleaned text. This preserves punctuation, emojis, formatting, and discourse markers that may signal communicative function. Thus, classical baselines receive normalized task-oriented text, while LLMs receive the closest available version of the public-facing post.

B LLM Prompts

B.1 Translation Prompt

We translated non-English Facebook posts into English using GPT-4.1-nano through a batched OpenAI translation pipeline with temperature set to 0. The user message contained a JSON list of records with `key`, `language`, and `text` fields.

Listing 1: System prompt used for translating Facebook posts into English.

```
You are a professional translation
    engine for public-sector social
    media posts.
Translate each text into clear English.

Rules:
- Preserve numbers, dates, street names,
     place names, hashtags, URLs, and
    emojis.
- Do not summarize.
- Do not add information.
- If the text is already English, return
     it unchanged.
- If text is empty, return an empty
    string.
- Return ONLY valid JSON as a list of
    objects:
  [{"key": "...", "translated_text":
      "..."}]
```

B.2 Zero-shot Classification Prompt

We used the following zero-shot prompt template for LLM classification. The model was asked to assign one label for Task A and one label for Task B, using only the predefined label set.

Listing 2: Zero-shot LLM classification prompt.

```
[SYSTEM MESSAGE]

You are a careful text classification
    assistant. Your task is to assign
    exactly one label for Task A and
    exactly one label for Task B.

Follow these rules:
- Use only the label IDs and names
    listed below.
- Choose the single best label for each
    task.
- Do not output explanations.
- Return valid JSON only.
- If the post is short or ambiguous,
    choose the most plausible label
    using the definitions below.

Task A labels:

A1 Water quality / safety / public
    health
Posts about water safety, drinking-water
     quality, contamination, testing,
    boil-water notices, bacteria,
    chemical risks, turbidity, chlorine,
     PFAS, or health risk.

A2 Service disruption / repair /
    infrastructure
Posts about outages, interruptions,
    repairs, leaks, bursts, maintenance,
     restoration, crews, pipelines,
    treatment plants, or infrastructure
    problems.
```

```
A3 Water conservation / drought / demand
    management
Posts about saving water, drought,
    restrictions, efficient use, demand
    reduction, conservation campaigns,
    or scarcity.

A4 Flood / stormwater / wastewater /
    sewer
Posts about flooding, heavy rain
    response, drains, sewer overflow,
    stormwater, wastewater incidents, or
    sewage.

A5 Environmental sustainability / waste
    / recycling / biodiversity
Posts about environmental protection,
    emissions, habitat, recycling,
    biodiversity, wetlands, river
    ecology, sustainability programs, or
    nature stewardship.

A6 Public education / heritage /
    community engagement
Posts about educational activities,
    awareness events, tours, school
    visits, museum/heritage content,
    public exhibitions, or community
    engagement not mainly about a
    current operational issue.

A7 Routine / institutional / customer
    service / other
Posts about office hours, billing, job
    vacancies, holiday notices, generic
    greetings, routine institutional
    communication, or items that do not
    fit A1-A6.

Task B labels:

B1 Alert / warning
Urgent alert or warning about a risk,
    disruption, incident, hazard, or
    emergency.

B2 Instruction / advice to public
Tells people what to do, for example
    boil water, avoid an area, conserve
    water, prepare, report leaks, or
    follow safety advice.

B3 Operational update / resolution
Status update about work underway,
    service restoration, incident
    handling, repair progress, or
    completion.

B4 Reassurance / safety information
Explains that water is safe, risks are
    low, monitoring is ongoing, or
    provides calming safety
    clarification.

B5 Education / awareness
Explains information for public
    understanding without mainly
    promoting the institution.

B6 Institutional promotion / community
    news
Promotes events, campaigns, achievements
    , awards, partnerships, or general
    community news.

B7 Other / unclear
Use when the communication function is
    too vague, ambiguous, context-
    dependent, or does not fit B1-B6.

Return this JSON schema exactly:
{
  "task_a": {
    "label_id": "...",
    "label_name": "..."
  },
  "task_b": {
    "label_id": "...",
    "label_name": "..."
  },
  "confidence": 0.00
}


[USER MESSAGE]

Classify this post:
{text}
```

B.3 Five-shot Classification Prompt

The five-shot condition used the same system instructions, label definitions, output schema, and classification rules as the zero-shot prompt in Listing 2. Five fixed demonstrations were then inserted before the final classification query, as shown below.

Listing 3: Five-shot LLM classification prompt with the five fixed demonstrations used in the experiments.

```


[USER MESSAGE]

Example 1
Text: Due to possible bacterial
    contamination, residents in the
    affected area must boil tap water
    before drinking or cooking until
    further notice.
Output:


[ASSISTANT MESSAGE]

{"task_a":{"label_id":"A1","label_name
    ":"Water quality / safety / public
    health"},"task_b":{"label_id":"B2","
    label_name":"Instruction / advice to
     public"},"confidence":0.98}


[USER MESSAGE]

Example 2
Text: Repair crews are working on a
    burst water pipe. Water pressure is
```

```
    low, and service is expected to
    return this evening.
Output:


[ASSISTANT MESSAGE]

{"task_a":{"label_id":"A2","label_name
    ":"Service disruption / repair /
    infrastructure"},"task_b":{"label_id
    ":"B3","label_name":"Operational
    update / resolution"},"confidence
    ":0.96}


[USER MESSAGE]

Example 3
Text: During the dry and hot period,
    please save water by avoiding garden
     watering during the day and using
    water wisely.
Output:


[ASSISTANT MESSAGE]

{"task_a":{"label_id":"A3","label_name
    ":"Water conservation / drought /
    demand management"},"task_b":{"
    label_id":"B2","label_name":"
    Instruction / advice to public"},"
    confidence":0.95}


[USER MESSAGE]

Example 4
Text: Heavy rain may overload drains and
     cause sewer overflow in low-lying
    areas. Avoid flooded streets and
    follow local warnings.
Output:


[ASSISTANT MESSAGE]

{"task_a":{"label_id":"A4","label_name
    ":"Flood / stormwater / wastewater /
     sewer"},"task_b":{"label_id":"B1","
    label_name":"Alert / warning"},"
    confidence":0.95}


[USER MESSAGE]

Example 5
Text: Learn why correct waste sorting
    and recycling protect rivers,
    wildlife, and local biodiversity.
Output:


[ASSISTANT MESSAGE]

{"task_a":{"label_id":"A5","label_name
    ":"Environmental sustainability /
    waste / recycling / biodiversity"},"
    task_b":{"label_id":"B5","label_name
    ":"Education / awareness"},"
    confidence":0.94}


[USER MESSAGE]

Now classify this new post:
{text}
```

B.4 Prompt Design Notes

The prompts were designed to make translation and classification outputs constrained, reproducible, and easy to parse. For translation, the system prompt framed the model as a public-sector social-media translation engine and instructed it to produce clear English while preserving numbers, dates, street names, place names, hashtags, URLs, and emojis; it also prohibited summarization and added information, and required a valid JSON list as output. For classification, each prompt defined the task context as Facebook communication by European water utilities, listed the complete Task A and Task B label sets, and required the model to choose exactly one label for each task. We kept the prompt structure fixed across models, languages, and shot settings so that differences reflected the model, text source, and number of examples rather than changes in prompt wording.

C Annotation

C.1 Inter-annotator agreement

Agreement was calculated on the independent annotations before discussion or adjudication. The most frequent disagreements involved adjacent operational/topic categories for Task A and distinctions among education, institutional communication, instruction, alert, and operational update for Task B.

Task	N	Raw agreement	Cohen's κ
Task A	1500	72.3%	0.672
Task B	1500	79.1%	0.715

Table 4: Inter-annotator agreement before consensus resolution.

C.2 Guidelines

Annotators assigned exactly one Task A label and one Task B label to each post. Task A represents the main topic or water-related issue discussed in the post, whereas Task B represents the primary communicative function of the post. The two tasks were labeled independently: the Task A decision

did not automatically determine the Task B decision.

Task B.1 was not annotated separately. It was derived after annotation by grouping the fine-grained Task B labels into four broader communicative-function categories.

C.3 General Annotation Procedure

Annotators followed the decision process below.

1. Read the complete translated post, including its title, sentence context, hashtags, and any visible formatting.
2. Consult the original-language text when the translation was unclear, incomplete, or appeared to mistranslate a place name, infrastructure term, warning, or instruction.
3. Identify the main subject of the post and assign one Task A label.
4. Identify what the utility is mainly trying to accomplish communicatively and assign one Task B label.
5. When several labels were possible, select the label that was most central to the complete post and most important for public water-utility communication.
6. Use the routine or unclear labels only when none of the more specific labels adequately described the post.

Labels were based on the overall meaning of the post rather than the presence of an isolated keyword. For example, the word *water* alone did not justify a water-quality label, and the word *warning* did not automatically indicate an alert if the post was describing a historical event or educational campaign.

C.4 Distinguishing Task A and Task B

Task A answers:

> *What is the post mainly about?*

Task B answers:

> *What is the utility mainly trying to do through this post?*

A post may therefore receive different but complementary labels across the two tasks. For example, a post about a pipe burst may be labeled as service disruption for Task A, while its Task B label may be:

- *Alert / warning*, when it announces the disruption;
- *Instruction / advice to public*, when it tells residents to store water;
- *Operational update / resolution*, when it reports repair progress; or
- *Reassurance / safety information*, when it confirms that service has been restored or the water is safe.

Annotators therefore considered the topic and communicative function separately rather than assigning Task B solely from the Task A label.

C.5 Task A: Main water-utility topic

A1: Water quality / safety / public health

Use A1 when the main subject is the safety or quality of drinking water, contamination, testing, monitoring, or a possible health risk.

Typical indicators include:

- boil-water or do-not-drink notices;
- bacterial, chemical, or biological contamination;
- PFAS, chlorine, turbidity, colour, taste, or odour;
- drinking-water sampling and laboratory results;
- statements about whether tap water is safe;
- health effects related to water consumption.

Examples:

> "Boil tap water before drinking because bacteria may be present."

> "Recent samples confirm that the water is safe to drink."

Do not use A1 when water quality is only mentioned incidentally and the main issue is an outage, pipe repair, flooding, wastewater operation, or general environmental education.

A2: Service disruption / repair / infrastructure

Use A2 when the main subject is water-service continuity, network operation, repair work, maintenance, construction, or infrastructure failure.

Typical indicators include:

- water outages or interruptions;
- low pressure or temporary loss of supply;
- pipe leaks, bursts, or network faults;
- planned maintenance or construction;
- repair crews and restoration work;
- pipelines, pumping stations, treatment facilities, or other infrastructure.

Examples:

> "Water supply will be interrupted tomorrow because of pipe maintenance."
>
> "Repair teams are working on a burst main causing low pressure."

Use A2 even when the post also includes an instruction such as storing water. The instruction affects Task B, while the underlying topic remains service disruption.

A3: Water conservation / drought / demand management

Use A3 when the main subject is water scarcity, drought, high demand, restrictions, or reducing water consumption.

Typical indicators include:

- drought or prolonged dry weather;
- water-use restrictions;
- high summer demand or low reservoir levels;
- requests to reduce consumption;
- efficient household or garden water use;
- conservation and demand-management campaigns.

Examples:

> "Please avoid watering gardens during the current dry period."
>
> "Small changes at home can reduce daily water demand."

A concrete request to save water is normally labeled B2 for Task B. A post that mainly explains why conservation matters without directing the public to act is normally labeled B5.

A4: Flood / stormwater / wastewater / sewer

Use A4 when the main subject is flooding, rainfall response, stormwater, drains, wastewater, sewers, sewage, or wastewater treatment.

Typical indicators include:

- heavy rain and local flooding;
- stormwater drainage;
- sewer overflow or blockage;
- sewage discharge or wastewater incidents;
- wastewater-treatment processes or facilities;
- runoff, drainage maintenance, or combined sewer systems.

Examples:

> "Heavy rain may overload drains and cause sewer overflow."
>
> "Wastewater-treatment operations have been affected by the storm."

Use A4 rather than A2 when the post primarily concerns sewer, wastewater, drainage, or flood management, even when infrastructure repair is also mentioned.

A5: Environmental sustainability / waste / recycling / biodiversity

Use A5 when the main subject is environmental protection, sustainability, waste management, recycling, biodiversity, ecosystems, or nature stewardship.

Typical indicators include:

- recycling and waste sorting;
- hazardous, plastic, or food waste;
- circular-economy initiatives;
- biodiversity, wetlands, rivers, or habitats;
- emissions, energy efficiency, or climate initiatives;
- environmental restoration and protection.

Examples:

> "Correct waste sorting helps prevent plastic from entering rivers."
>
> "The new wetland project supports local biodiversity."

Use A5 when the environmental issue is the subject of the message. Use A6 when the post is mainly about a school visit, exhibition, tour, or community activity, even if the event includes environmental content.

A6: Public education / heritage / community engagement

Use A6 when the main subject is an educational, cultural, historical, or community-oriented activity related to water.

Typical indicators include:

- school visits and workshops;
- museum or treatment-plant tours;
- exhibitions and public-information events;
- water heritage and historical infrastructure;
- World Water Day activities;
- community participation not centred on a current incident.

Examples:

> "Students visited the water museum to learn about the water cycle."
>
> "Join a guided tour of the city's historic water tower."

Use A6 when the activity itself is the central topic. Use A5 when the main subject is an environmental programme, ecosystem, sustainability policy, or waste-management issue.

A7: Routine / institutional / customer service / other

Use A7 for routine organizational communication or posts whose topic does not fit A1-A6.

Typical indicators include:

- billing, payment, or customer-service information;
- office hours and contact details;
- recruitment and job vacancies;
- holiday greetings or closure notices;
- awards, internal achievements, or staff news;
- general institutional announcements;
- content unrelated to a specific water-risk category.

Examples:

> "Our customer-service office will be closed during the holiday."
>
> "We are recruiting a new project manager."

A7 is the fallback Task A category, but it should not be used when a specific operational, public-health, drought, wastewater, environmental, or educational topic is clearly present.

C.6 Task B: Primary communicative function

B1: Alert / warning

Use B1 when the primary function is to notify the public of an urgent or potentially harmful situation.

Typical indicators include:

- warnings about contamination, outages, flooding, or hazards;
- announcements of an unexpected incident;
- urgent notices using language such as *warning*, *attention*, *emergency*, or *until further notice*;
- information whose main purpose is to make the public aware that a risk or disruption exists.

Examples:

> "Warning: water supply has been interrupted in the northern district."
>
> "Heavy rain may cause flooding in low-lying areas."

When a post both announces an incident and gives a clear instruction, use B2 if the requested action is the dominant communicative purpose. Use B1 when notifying the public of the risk is primary and any advice is secondary.

B2: Instruction / advice to public

Use B2 when the main function is to tell people what action to take.

Typical indicators include:

- boil, avoid, store, conserve, report, prepare, or evacuate;
- safety precautions;
- requests to reduce water use;
- instructions for reporting leaks;
- waste-sorting or disposal directions;
- steps to follow before, during, or after an incident.

Examples:

> "Boil water for at least one minute before drinking."
>
> "Please store enough water before tomorrow's interruption."

General educational tips may also contain imperative wording. Use B2 when the post asks the public to perform a specific action. Use B5 when the main purpose is explanation or awareness rather than immediate behaviour.

B3: Operational update / resolution

Use B3 when the primary function is to report the status, progress, or resolution of operational work or an incident.

Typical indicators include:

- repair work is underway;
- crews are on site;
- expected restoration times;
- changes in service status;
- confirmation that a fault has been repaired;
- confirmation that service has returned.

Examples:

> "Repairs are continuing, and service is expected to return by 18:00."
>
> "The pipe fault has been repaired and water service is restored."

Use B3 when the post primarily reports what the utility is doing or what stage the incident has reached. An initial announcement of a new problem is normally B1, while a request for public action is normally B2.

B4: Reassurance / safety information

Use B4 when the primary function is to reduce concern or clarify that a risk is controlled, limited, or absent.

Typical indicators include:

- confirmation that water is safe;
- negative contamination results;
- statements that monitoring is continuing;
- explanations that unusual taste, colour, or smell is harmless;
- statements intended to calm concern or maintain confidence.

Examples:

> "Tests found no contamination, and the water is safe to drink."
>
> "The unusual taste is temporary and does not present a health risk."

Use B4 when safety clarification or reassurance is the main purpose. Use B3 when the main emphasis is repair progress or restoration of service.

B5: Education / awareness

Use B5 when the post mainly explains a topic, provides general knowledge, or raises awareness.

Typical indicators include:

- explanations of water treatment or wastewater processes;
- facts about conservation, recycling, climate, or water quality;
- educational campaigns;
- general advice presented primarily as awareness material;
- informational content without a current incident or direct operational request.

Examples:

"Learn how wastewater is treated before being returned to the environment."

"This campaign explains why recycling protects waterways."

Use B5 when teaching or awareness is the main purpose. Use B6 when the main purpose is promoting an event, partnership, achievement, or the utility itself.

B6: Institutional promotion / community news

Use B6 when the main function is to promote the utility, an event, campaign, partnership, achievement, or general community activity.

Typical indicators include:

- event invitations;
- awards and achievements;
- partnerships and sponsorships;
- recruitment or staff announcements;
- celebrations and holiday greetings;
- community news and organizational promotion.

Examples:

"Visit our information stand at the city festival this weekend."

"We are proud to announce a new partnership with the local university."

A campaign may contain educational material. Use B6 when promoting the campaign, organization, or event is dominant; use B5 when explaining the substantive issue is dominant.

B7: Other / unclear

Use B7 only when the communicative function cannot be assigned confidently.

Appropriate cases include:

- extremely short or context-free messages;
- unclear or incomplete translations;
- posts consisting mainly of a slogan, image reference, or external link;
- posts with no identifiable warning, instruction, update, reassurance, educational, or promotional function.

Examples:

"Water is life."

"More information coming soon."

Annotators were instructed not to use B7 merely because a post contained several functions. In such cases, they selected the dominant function using the rules below.

C.7 Rules for Ambiguous and Multi-Function Posts

Topic plus instruction. When a post describes an issue and tells the public what to do, Task A captures the issue and Task B normally captures the instruction.

"A pipe burst has interrupted supply. Please store water for the evening."

Task A: A2; Task B: B2.

Incident announcement versus operational update. Use B1 when the post mainly announces a new or ongoing risk. Use B3 when it mainly reports repair progress, expected restoration, or resolution.

Instruction versus education. Use B2 when the public is asked to perform a clear action. Use B5 when the post mainly explains why an action or issue matters.

Operational update versus reassurance. Use B3 when the focus is the status of work or service. Use B4 when the focus is safety, risk reduction, or calming public concern.

Education versus institutional promotion. Use B5 when the substantive information is central. Use B6 when the main purpose is to advertise an event, campaign, partnership, or organizational achievement.

Current event versus historical or educational content. Words such as *flood*, *drought*, or *contamination* were interpreted in context. A museum post about a historical flood, for example, could be labeled A6 rather than A4 when the main purpose was heritage or public education.

Multiple operational topics. When several Task A categories were present, annotators selected the topic that motivated the post or created the most important water-related consequence. For example, a pipe failure causing temporary discoloured water

could be A2 if the post mainly concerned repair and service interruption, but A1 if the central message was a drinking-water safety warning.

Hashtags and links. Hashtags were treated as supporting context rather than decisive evidence. External links were not assumed to determine the label unless the surrounding post text clearly described their purpose.

Dominant-function principle. When a post contained several communicative functions, annotators selected the function receiving the greatest emphasis, appearing in the main statement, or carrying the most important public-risk implication. The presence of several functions alone did not justify B7.

C.8 Derived Grouped Task B.1

Task B.1 was generated deterministically from the final Task B labels:

- B1, B2, B3, and B4 were grouped as *Risk / incident communication*;
- B5 remained *Education / awareness*;
- B6 became *Institutional / community communication*;
- B7 remained *Other / unclear*.

This grouping was introduced because real utility posts frequently combine warnings, instructions, operational updates, and reassurance. Task B retains these distinctions for fine-grained evaluation, while Task B.1 evaluates the broader communicative role.

D Utility Accounts

Table 5 summarizes the public Facebook utility accounts included in the corpus. For each account, we report the assigned utility identifier, inferred utility or account name, associated city or region, number of posts, dominant language, and temporal coverage. The table shows that the corpus covers multiple European cities and languages, but is unevenly distributed across utilities, reflecting differences in Facebook activity and available timeline length.

E Baseline Experiment Settings

Table 6 summarizes the main baseline settings used in the supervised and rule-based experiments. These settings document the text variants, model parameters, and evaluation protocol used for the non-LLM baselines.

F LLM configurations

Table 7 summarizes the model identifiers, inference backends, decoding settings, and evaluation conditions used for the LLM experiments. For the local Ollama runs, the exact runtime build and quantization metadata were not preserved. We therefore report the model identifiers used at inference rather than inferring unrecorded implementation details.

G Sampling and Keyword-Rule Lexicons

This appendix reports the complete English lexicons used for two distinct purposes: keyword-guided sampling of the annotation set and the deterministic keyword-rule classification baseline. The sampling lexicon was used only to increase coverage of domain-relevant posts and did not determine gold labels. Tables 8–10 report the sampling and baseline lexicons.

Because the sampling lexicon and keyword baseline were derived from the same domain taxonomy and share lexical cues, the keyword baseline should be interpreted as an in-domain lexical reference rather than a fully independent baseline.

Utility ID	Inferred utility or account	City/region	Posts	Dominant language	Coverage
U03	EPAL Águas de Lisboa	Lisbon	4,973	Portuguese	2011-2026
U01	HSY	Helsinki region	3,017	Finnish	2014-2026
U08	Eau de Paris	Paris	2,476	French	2009-2026
U09	ABC Napoli	Naples	2,164	Italian	2017-2026
U05	Berliner Wasserbetriebe	Berlin	1,550	German	2014-2026
U10	Stockholm Vatten och Avfall	Stockholm	1,398	Swedish	2009-2026
U02	Hamburg Wasser	Hamburg	1,243	German	2018-2026
U04	Oslo Vann	Oslo	1,062	Norwegian	2013-2026
U12	Turun Vesihuolto	Turku	744	Finnish	2016-2026
U07	Farys	Ghent	681	Dutch	2014-2026
U11	Tampereen Vesi	Tampere	466	Finnish	2020-2026
U06	World Waternet	Amsterdam	226	English	2023-2026

Table 5: Utility accounts included in the Facebook corpus. Post counts refer to text-bearing posts used in the dataset after preprocessing. Coverage indicates the year range represented for each account.

Baseline	Setting	Value
All trainable baselines	Evaluation protocol	Stratified five-fold cross-validation on the 1,500-post gold-standard sample
All trainable baselines	Fold construction	`StratifiedKFold(n_splits=5, shuffle=True, random_state=42)`
All trainable baselines	Score aggregation	Predictions from the five held-out folds are pooled so that each gold-standard post is evaluated exactly once by a model that was not trained on that post
All trainable baselines	Primary reported scores	Metrics computed from the pooled out-of-fold predictions; fold-level means and standard deviations are also retained for assessing variability
All trainable baselines	Random seed	42
Majority baseline	Prediction rule	Predicts the most frequent label in the training portion of each fold
Majority baseline	Evaluation	Applied independently within each of the five cross-validation folds and evaluated using pooled out-of-fold predictions
Keyword rules	Text input	`translated_clean`
Keyword rules	Matching rule	Assigns the first label whose manually curated lexical triggers match the post text
Keyword rules	Default label, Task A	*Routine / institutional / customer service / other*
Keyword rules	Default label, Task B	*Other / unclear*
Keyword rules	Evaluation	Fixed rules require no training and are evaluated on the held-out portion of each fold; predictions are pooled across all five folds
TF-IDF + LR	Text variants	`native_clean` and `translated_clean`
TF-IDF + LR	Features	Word unigram and bigram TF-IDF features, `ngram_range=(1,2)`
TF-IDF + LR	Maximum features	30,000
TF-IDF + LR	Minimum document frequency	2
TF-IDF + LR	Classifier	Logistic regression
TF-IDF + LR	Logistic-regression parameters	`max_iter=2000`, `class_weight=balanced`, and `random_state=42`
TF-IDF + LR	Fold fitting	The TF-IDF vectorizer and logistic-regression classifier are fitted only on the training portion of each fold
Multilingual embeddings + LR	Text variants	`native_clean` and `translated_clean`
Multilingual embeddings + LR	Encoder	`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`
Multilingual embeddings + LR	Embedding settings	`batch_size=64` and `normalize_embeddings=True`
Multilingual embeddings + LR	Classifier	Logistic regression with `max_iter=2000`, `class_weight=balanced`, and `random_state=42`
Multilingual embeddings + LR	Fold fitting	The sentence encoder remains fixed; a new logistic-regression classifier is fitted on the training embeddings in each fold
Task B.1	Evaluation	Task B.1 predictions are obtained post hoc by mapping the Task B gold labels and out-of-fold predictions to the four grouped communication functions; no separate Task B.1 classifier is trained

Table 6: Baseline experiment settings for the majority, keyword-rule, TF-IDF logistic-regression, and multilingual sentence-embedding classifiers. Trainable baselines are evaluated using stratified five-fold cross-validation, and their primary scores are computed from pooled out-of-fold predictions.

Model	Identifier, size, and backend	Inference settings	Evaluation conditions
Gemma 4	`gemma4:e4b`; E4B effective-parameter variant; Ollama	Temperature = 0; JSON output; reasoning disabled (`think=false`); context window = 4096; max. generation = 256 tokens	Native and English-translated; zero-shot and five-shot
Qwen 3.5	`qwen3.5:9b`; 9B parameters; Ollama	Temperature = 0; JSON output; reasoning disabled (`think=false`); context window = 4096; max. generation = 256 tokens	Native and English-translated; zero-shot and five-shot
GPT-4.1-mini	`gpt-4.1-mini`; size not publicly disclosed; OpenAI API	Temperature = 0; JSON-object output	Native and English-translated; zero-shot and five-shot
GPT-4o-mini	`gpt-4o-mini`; size not publicly disclosed; OpenAI API	Temperature = 0; JSON-object output	Native and English-translated; zero-shot and five-shot

Table 7: LLM configurations used for classification. Each configuration predicts one Task A label and one Task B label for each of the 1,500 gold-standard posts. Task B.1 is derived post hoc by mapping Task B predictions to the grouped label space.

Sampling group	Keywords and phrases
Water quality / safety / contamination	water quality; drinking water; tap water; safe to drink; water safety; contamination; contaminated; pollution; turbidity; chlorine; bacteria; e coli; ecoli; coliform; legionella; pfas; boil water; boil; water sampling; water monitoring; quality monitoring; public health; health risk; beach quality; water test; water testing; water analysis
Service disruption / repair / infrastructure	water interruption; service interruption; water service interruption; interruption; disruption; outage; no water; water cut; low pressure; pressure; pipe burst; burst pipe; broken pipe; leak; fault; repair; maintenance; construction; water network; network; restoration time; water supply interruption; service alert; planned maintenance; emergency repair
Flood / stormwater / wastewater / sewer	flood; flooding; storm; stormwater; heavy rain; rainfall; rainwater; runoff; drainage; drain; sewer; sewerage; wastewater; wastewater treatment; overflow; combined sewer overflow; sewage; treatment plant
Water conservation / drought / demand management	drought; scarcity; water scarcity; water restriction; water restrictions; water conservation; water saving; save water; conserve water; water consumption; water demand; heatwave; heat wave; dry spell; sustainable water use; conservation tips; watering ban; water wisely
Environmental sustainability / waste / recycling / biodiversity	waste sorting; waste management; waste collection; recycling; hazardous waste; plastic waste; food waste; biowaste; bio waste; battery; circular economy; biodiversity; wildlife; ecosystem; baltic sea; air quality; environmental conservation; sustainability; climate change; invasive species; compost; organic waste
Education / awareness / campaign	education; educational; awareness; campaign; outreach; school; children; student; students; workshop; training; tour; museum; exhibition; world water day; tips; public awareness; learn; learning; webinar
Routine / institutional / community	holiday; christmas; greeting; celebration; anniversary; award; event; community engagement; community events; partnership; customer service; billing; bill; payment; job; recruitment; career; contest; cultural; heritage; promotion; brand; opening hours; office closed
Rare/crisis prioritization signal	boil; contaminat; bacteria; e coli; ecoli; legionella; pfas; outbreak; illness; turbidity; overflow; flood; drought; restriction; outage; interruption; pipe burst; burst pipe; broken pipe; low pressure; no water; sewer; wastewater

Table 8: English keyword groups used for constructing the keyword-guided annotation sample. Posts were matched on cleaned English-translated text using case-insensitive matching. If a post matched several groups, the first group in the predefined priority order was assigned. The final row lists terms used to prioritize rare or crisis-related candidates within each sampling group.

Task A label	Keyword rules
A1: Water quality / safety / public health	water quality; drinking water; tap water; safe to drink; water safety; contamination; contaminated; pollution; turbidity; chlorine; bacteria; e coli; ecoli; coliform; legionella; pfas; boil water; boil; water sampling; water monitoring; quality monitoring; public health; health risk; water test; water testing
A2: Service disruption / repair / infrastructure	water interruption; service interruption; interruption; disruption; outage; no water; water cut; low pressure; pipe burst; burst pipe; broken pipe; leak; fault; repair; maintenance; construction; water network; restoration time; planned maintenance; emergency repair
A3: Water conservation / drought / demand management	drought; scarcity; water scarcity; water restriction; water restrictions; water conservation; water saving; save water; conserve water; water consumption; water demand; heatwave; heat wave; dry spell; watering ban
A4: Flood / stormwater / wastewater / sewer	flood; flooding; stormwater; storm water; heavy rain; rainfall; runoff; drainage; sewer; sewerage; wastewater; wastewater treatment; overflow; sewage; treatment plant
A5: Environmental sustainability / waste / recycling / biodiversity	waste sorting; waste management; waste collection; recycling; hazardous waste; plastic waste; food waste; biowaste; battery; circular economy; biodiversity; wildlife; ecosystem; air quality; environmental conservation; sustainability; climate change; invasive species; compost
A6: Public education / heritage / community engagement	education; educational; awareness; campaign; outreach; school; children; student; workshop; training; tour; museum; exhibition; world water day; public awareness; heritage; history
A7: Routine / institutional / customer service / other	holiday; christmas; greeting; celebration; anniversary; award; event; partnership; customer service; billing; bill; payment; job; recruitment; career; contest; opening hours; office closed

Table 9: Task A keyword rules used by the translated-text keyword baseline. Rules were evaluated in label order from A1 to A7 using case-insensitive substring matching. The first matching label was returned; posts with no match defaulted to A7 (*Routine / institutional / customer service / other*).

Task B label	Keyword rules
B1: Alert / warning	warning; alert; urgent; emergency; do not drink; do not use; risk; danger; unsafe; interruption; outage; flood warning; service alert
B2: Instruction / advice to public	please; boil; boil water; save water; conserve water; avoid; do not; report; contact; prepare; store water; follow; use less; sort; recycle; bring; check
B3: Operational update / resolution	update; updated; ongoing; resolved; fixed; restored; completed; repair work; restoration; back to normal; service has returned; work continues
B4: Reassurance / safety information	safe; safe to drink; no risk; no contamination; monitoring; tested; test results; under control; normal; quality is good
B5: Education / awareness	learn; education; educational; awareness; did you know; tips; guide; explains; how; why; campaign; workshop; school; students
B6: Institutional promotion / community news	event; join us; celebration; holiday; award; proud; partnership; recruitment; job; career; museum; exhibition; festival; community
B7: Other / unclear	No explicit keywords. Used as the fallback when none of the B1-B6 rules matched.

Table 10: Task B keyword rules used by the translated-text keyword baseline. Rules were evaluated in label order from B1 to B6 using case-insensitive substring matching. The first matching label was returned; posts with no match defaulted to B7 (*Other / unclear*).

H Additional Evaluation Metrics

H.1 Overall accuracy and weighted-F1

The main paper reports macro-F1 as the primary metric because the gold-standard label distributions are imbalanced. For completeness, Tables 11 and 12 report overall accuracy and weighted-F1.

H.2 Five-fold stability of supervised baselines

Table 13 reports the mean and sample standard deviation of macro-F1 across the five held-out folds for each supervised baseline. These values complement the pooled out-of-fold scores reported in the main paper by showing fold-to-fold variability.

H.3 Per-language native-translated analysis

To examine RQ2 by language, we report results for languages represented by at least 50 posts in the gold set: Portuguese, Finnish, Italian, German, Swedish, Norwegian, French, and Dutch. English ($n = 5$) and Spanish ($n = 1$) are omitted from separate performance estimates because their samples are too small. Danish and records with missing language metadata had no examples in the gold set. Within each language, macro-F1 is computed over labels represented in that language subset. These analyses are descriptive because sample sizes and class coverage differ substantially across languages. Tables 14 and 15 report the corresponding translated–native differences for Tasks A and B.

H.4 Bootstrap confidence intervals

Tables 16 and 17 report uncertainty estimates from 10,000 stratified-bootstrap replicates with random seed 42. Individual intervals are 95% percentile bootstrap confidence intervals. Paired differences use identical resampled post indices for the native and translated conditions.

Model	Text	Setting	Task A	Task B	Task B.1	Invalid Rows %
Majority	all	Stratified 5-fold CV	0.316	0.425	0.425	–
Keyword rules	translated	rule-based	0.562	0.387	0.422	–
TF-IDF + LR	translated	Stratified 5-fold CV	0.685	0.717	0.728	–
TF-IDF + LR	native	Stratified 5-fold CV	0.555	0.622	0.633	–
Multilingual embeddings + LR	translated	Stratified 5-fold CV	0.669	0.665	0.685	–
Multilingual embeddings + LR	native	Stratified 5-fold CV	0.662	0.643	0.667	–
Gemma 4	translated	zero-shot	0.646	0.552	0.689	0.00
Gemma 4	translated	five-shot	0.579	0.434	0.576	12.27
Gemma 4	native	zero-shot	0.626	0.535	0.674	0.00
Gemma 4	native	five-shot	0.532	0.399	0.527	19.80
GPT-4.1-mini	translated	zero-shot	0.653	0.501	0.612	0.00
GPT-4.1-mini	translated	five-shot	0.647	0.482	0.595	0.00
GPT-4.1-mini	native	zero-shot	0.635	0.481	0.604	0.07
GPT-4.1-mini	native	five-shot	0.631	0.470	0.589	0.00
GPT-4o-mini	translated	zero-shot	0.618	0.601	0.742	0.00
GPT-4o-mini	translated	five-shot	0.643	0.588	0.723	0.00
GPT-4o-mini	native	zero-shot	0.602	0.584	0.728	0.00
GPT-4o-mini	native	five-shot	0.623	0.575	0.714	0.00
Qwen 3.5	translated	zero-shot	0.606	0.458	0.603	0.13
Qwen 3.5	translated	five-shot	0.612	0.484	0.616	0.13
Qwen 3.5	native	zero-shot	0.587	0.448	0.589	0.53
Qwen 3.5	native	five-shot	0.592	0.454	0.600	1.20

Table 11: Overall accuracy on the 1,500-post gold standard. Supervised baseline scores are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. Keyword rules and LLMs are evaluated on all 1,500 annotated posts. Invalid Rows % denotes the percentage of LLM outputs with a missing, malformed, or out-of-schema Task A or Task B label.

Model	Text	Setting	Task A	Task B	Task B.1	Invalid Rows %
Majority	all	Stratified 5-fold CV	0.152	0.254	0.254	–
Keyword rules	translated	rule-based	0.576	0.414	0.426	–
TF-IDF + LR	translated	Stratified 5-fold CV	0.690	0.717	0.727	–
TF-IDF + LR	native	Stratified 5-fold CV	0.558	0.626	0.630	–
Multilingual embeddings + LR	translated	Stratified 5-fold CV	0.673	0.682	0.686	–
Multilingual embeddings + LR	native	Stratified 5-fold CV	0.667	0.663	0.665	–
Gemma 4	translated	zero-shot	0.641	0.547	0.674	0.00
Gemma 4	translated	five-shot	0.609	0.442	0.588	12.27
Gemma 4	native	zero-shot	0.625	0.536	0.662	0.00
Gemma 4	native	five-shot	0.582	0.417	0.558	19.80
GPT-4.1-mini	translated	zero-shot	0.647	0.511	0.617	0.00
GPT-4.1-mini	translated	five-shot	0.645	0.463	0.576	0.00
GPT-4.1-mini	native	zero-shot	0.630	0.485	0.605	0.07
GPT-4.1-mini	native	five-shot	0.628	0.447	0.565	0.00
GPT-4o-mini	translated	zero-shot	0.608	0.602	0.741	0.00
GPT-4o-mini	translated	five-shot	0.639	0.556	0.710	0.00
GPT-4o-mini	native	zero-shot	0.587	0.583	0.725	0.00
GPT-4o-mini	native	five-shot	0.617	0.547	0.703	0.00
Qwen 3.5	translated	zero-shot	0.603	0.447	0.595	0.13
Qwen 3.5	translated	five-shot	0.609	0.431	0.567	0.13
Qwen 3.5	native	zero-shot	0.585	0.440	0.584	0.53
Qwen 3.5	native	five-shot	0.591	0.404	0.564	1.20

Table 12: Overall weighted-F1 on the 1,500-post gold standard. Supervised baseline scores are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. Keyword rules and LLMs are evaluated on all 1,500 annotated posts. Invalid Rows % denotes the percentage of LLM outputs with a missing, malformed, or out-of-schema Task A or Task B label.

Model	Text	Task A	Task B	Task B.1
Majority	all	0.069 ± 0.000	0.085 ± 0.000	0.149 ± 0.000
Keyword rules	translated	0.563 ± 0.021	0.307 ± 0.021	0.356 ± 0.025
TF-IDF + LR	translated	0.672 ± 0.028	0.464 ± 0.054	0.548 ± 0.012
TF-IDF + LR	native	0.520 ± 0.039	0.395 ± 0.029	0.475 ± 0.024
Multilingual embeddings + LR	translated	0.656 ± 0.030	0.502 ± 0.053	0.587 ± 0.052
Multilingual embeddings + LR	native	0.650 ± 0.038	0.492 ± 0.046	0.571 ± 0.043

Table 13: Mean macro-F1 and sample standard deviation across the five held-out folds. Each fold contains 300 evaluation posts. These values describe fold-to-fold variability; the main model-comparison table reports macro-F1 computed from the pooled out-of-fold predictions over all 1,500 posts.

	Gemma 4		GPT-4.1-mini		GPT-4o-mini		Qwen 3.5	
Language (n)	0s	5s	0s	5s	0s	5s	0s	5s
Portuguese (420)	+0.009	+0.040	+0.012	+0.022	-0.006	+0.022	-0.028	+0.012
Finnish (363)	+0.008	+0.086	+0.003	-0.020	+0.018	+0.048	+0.022	+0.020
Italian (201)	+0.151	+0.029	+0.073	+0.196	+0.010	-0.067	+0.135	+0.063
German (191)	+0.062	0.000	+0.042	+0.073	+0.039	+0.006	+0.078	+0.049
Swedish (97)	+0.096	+0.102	+0.120	+0.134	-0.009	-0.016	+0.033	+0.056
Norwegian (83)	+0.054	+0.005	+0.013	+0.068	+0.004	-0.024	-0.011	-0.058
French (83)	+0.061	+0.005	-0.056	-0.054	+0.150	-0.095	-0.061	-0.016
Dutch (56)	+0.077	-0.098	+0.102	+0.008	+0.004	+0.045	+0.161	-0.037

Table 14: Per-language translated–native differences in Task A macro-F1. Values are $\Delta = \mathrm{F1}_{\mathrm{translated}} - \mathrm{F1}_{\mathrm{native}}$, so positive values indicate higher performance on English-translated input. 0s and 5s denote zero-shot and five-shot prompting. Results are shown for languages with at least 50 posts in the gold set.

	Gemma 4		GPT-4.1-mini		GPT-4o-mini		Qwen 3.5	
Language (n)	0s	5s	0s	5s	0s	5s	0s	5s
Portuguese (420)	-0.010	+0.022	-0.002	+0.005	+0.015	-0.038	-0.017	+0.018
Finnish (363)	-0.024	+0.022	+0.181	+0.096	+0.052	-0.024	+0.012	-0.013
Italian (201)	+0.030	-0.002	+0.111	+0.058	+0.019	+0.047	-0.011	-0.004
German (191)	+0.046	+0.154	+0.104	+0.169	+0.073	+0.028	+0.022	+0.046
Swedish (97)	+0.005	+0.116	+0.017	+0.010	+0.024	+0.012	-0.001	+0.017
Norwegian (83)	+0.018	+0.084	-0.009	-0.004	+0.009	+0.013	+0.013	+0.012
French (83)	+0.230	+0.064	+0.016	+0.004	-0.173	+0.054	-0.019	-0.101
Dutch (56)	+0.011	-0.044	-0.020	-0.007	+0.072	+0.072	+0.018	+0.053

Table 15: Per-language translated–native differences in Task B macro-F1. Values are $\Delta = \mathrm{F1}_{\mathrm{translated}} - \mathrm{F1}_{\mathrm{native}}$, so positive values indicate higher performance on English-translated input. 0s and 5s denote zero-shot and five-shot prompting. Results are shown for languages with at least 50 posts in the gold set.

Model	Text	Setting	Task A	Task B	Task B.1
Gemma 4	translated	0-shot	0.658 [0.634, 0.682]	0.352 [0.319, 0.388]	0.563 [0.522, 0.605]
Gemma 4	translated	5-shot	0.624 [0.597, 0.650]	0.331 [0.293, 0.369]	0.553 [0.505, 0.602]
Gemma 4	native	0-shot	0.633 [0.606, 0.658]	0.346 [0.314, 0.380]	0.552 [0.519, 0.585]
Gemma 4	native	5-shot	0.566 [0.536, 0.593]	0.292 [0.252, 0.332]	0.497 [0.459, 0.537]
GPT-4.1-mini	translated	0-shot	0.638 [0.609, 0.666]	0.359 [0.320, 0.399]	0.494 [0.472, 0.515]
GPT-4.1-mini	translated	5-shot	0.617 [0.586, 0.646]	0.325 [0.289, 0.362]	0.466 [0.443, 0.488]
GPT-4.1-mini	native	0-shot	0.616 [0.588, 0.644]	0.328 [0.293, 0.363]	0.491 [0.469, 0.512]
GPT-4.1-mini	native	5-shot	0.596 [0.566, 0.626]	0.315 [0.281, 0.351]	0.463 [0.441, 0.484]
GPT-4o-mini	translated	0-shot	0.620 [0.593, 0.646]	0.392 [0.346, 0.437]	0.646 [0.591, 0.698]
GPT-4o-mini	translated	5-shot	0.624 [0.595, 0.651]	0.381 [0.331, 0.429]	0.660 [0.601, 0.714]
GPT-4o-mini	native	0-shot	0.601 [0.573, 0.627]	0.384 [0.338, 0.428]	0.623 [0.565, 0.677]
GPT-4o-mini	native	5-shot	0.605 [0.576, 0.633]	0.372 [0.326, 0.418]	0.650 [0.603, 0.697]
Qwen 3.5	translated	0-shot	0.567 [0.538, 0.595]	0.277 [0.244, 0.313]	0.472 [0.448, 0.496]
Qwen 3.5	translated	5-shot	0.584 [0.554, 0.612]	0.269 [0.235, 0.307]	0.441 [0.413, 0.474]
Qwen 3.5	native	0-shot	0.558 [0.528, 0.587]	0.277 [0.242, 0.315]	0.461 [0.438, 0.486]
Qwen 3.5	native	5-shot	0.573 [0.543, 0.601]	0.242 [0.209, 0.277]	0.430 [0.407, 0.458]

Table 16: LLM macro-F1 with 95% percentile bootstrap confidence intervals from 10,000 stratified-bootstrap replicates. Invalid predictions are retained and scored as incorrect.

Model	Setting	Task A	Task B	Task B.1
Gemma 4	0-shot	+0.025 [+0.007, +0.045]	+0.007 [-0.014, +0.026]	+0.012 [-0.017, +0.038]
Gemma 4	5-shot	+0.058 [+0.033, +0.085]	+0.039 [+0.004, +0.076]	+0.056 [+0.019, +0.092]
GPT-4.1-mini	0-shot	+0.022 [+0.002, +0.043]	+0.032 [+0.010, +0.054]	+0.003 [-0.015, +0.021]
GPT-4.1-mini	5-shot	+0.021 [-0.002, +0.043]	+0.009 [-0.013, +0.033]	+0.003 [-0.016, +0.021]
GPT-4o-mini	0-shot	+0.019 [-0.001, +0.039]	+0.008 [-0.025, +0.039]	+0.023 [-0.004, +0.058]
GPT-4o-mini	5-shot	+0.018 [-0.002, +0.038]	+0.009 [-0.022, +0.037]	+0.010 [-0.034, +0.049]
Qwen 3.5	0-shot	+0.009 [-0.016, +0.035]	+0.000 [-0.019, +0.020]	+0.010 [-0.013, +0.035]
Qwen 3.5	5-shot	+0.011 [-0.012, +0.034]	+0.027 [-0.001, +0.052]	+0.011 [-0.009, +0.036]

Table 17: Paired translated–native differences in LLM macro-F1. Values are $\Delta = \mathrm{F1}_{\mathrm{translated}} - \mathrm{F1}_{\mathrm{native}}$ with 95% percentile bootstrap confidence intervals from 10,000 stratified-bootstrap replicates. Positive values favor translated input. Intervals excluding zero indicate evidence for a difference under this bootstrap analysis.

H.5 Classwise performance

Figures 4-9 show substantial variation across classes, while Table 18 gives the key used for each label. On Task A, most systems perform particularly strongly on service disruption and infrastructure posts, while water-quality posts remain more difficult. On Task B, performance is considerably lower for rare classes such as operational resolution, reassurance, and other or unclear communication. Grouping the overlapping communication functions in Task B.1 produces more stable classwise performance, particularly for risk and incident communication.

H.6 Detailed classwise results

Tables 19-24 provide precision, recall, F1, and support for the strongest supervised baseline and strongest LLM configuration on each task, complementing the classwise visualizations above.

Task	Code	Class label
Task A	A1	Water quality / safety / public health
Task A	A2	Service disruption / repair / infrastructure
Task A	A3	Water conservation / drought / demand management
Task A	A4	Flood / stormwater / wastewater / sewer
Task A	A5	Environmental sustainability / waste / recycling / biodiversity
Task A	A6	Public education / heritage / community engagement
Task A	A7	Routine / institutional / customer service / other
Task B	B1	Alert / warning
Task B	B2	Instruction / advice to public
Task B	B3	Operational update / resolution
Task B	B4	Reassurance / safety information
Task B	B5	Education / awareness
Task B	B6	Institutional promotion / community news
Task B	B7	Other / unclear
Task B.1	G1	Risk / incident communication
Task B.1	G2	Education / awareness
Task B.1	G3	Institutional / community communication
Task B.1	G4	Other / unclear

Table 18: Class codes used in the classwise evaluation figures and supplementary tables.

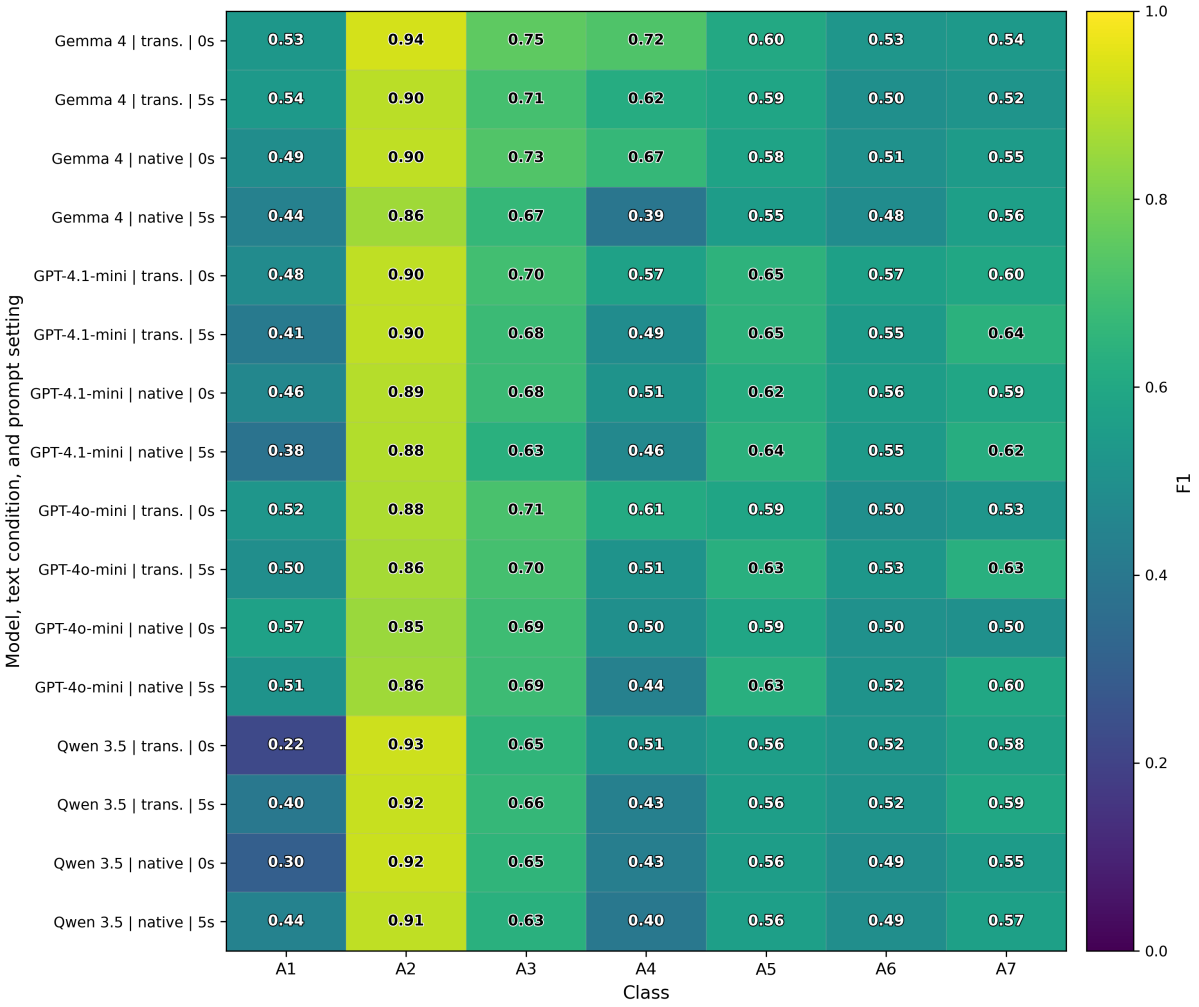


Figure 4: Classwise F1 for LLMs on Task A. Values are computed on all 1,500 gold-standard posts. A1-A7 denote the Task A labels defined in Table 18.



Figure 5: Classwise F1 for LLMs on Task B. Values are computed on all 1,500 gold-standard posts. B1-B7 denote the Task B labels defined in Table 18.

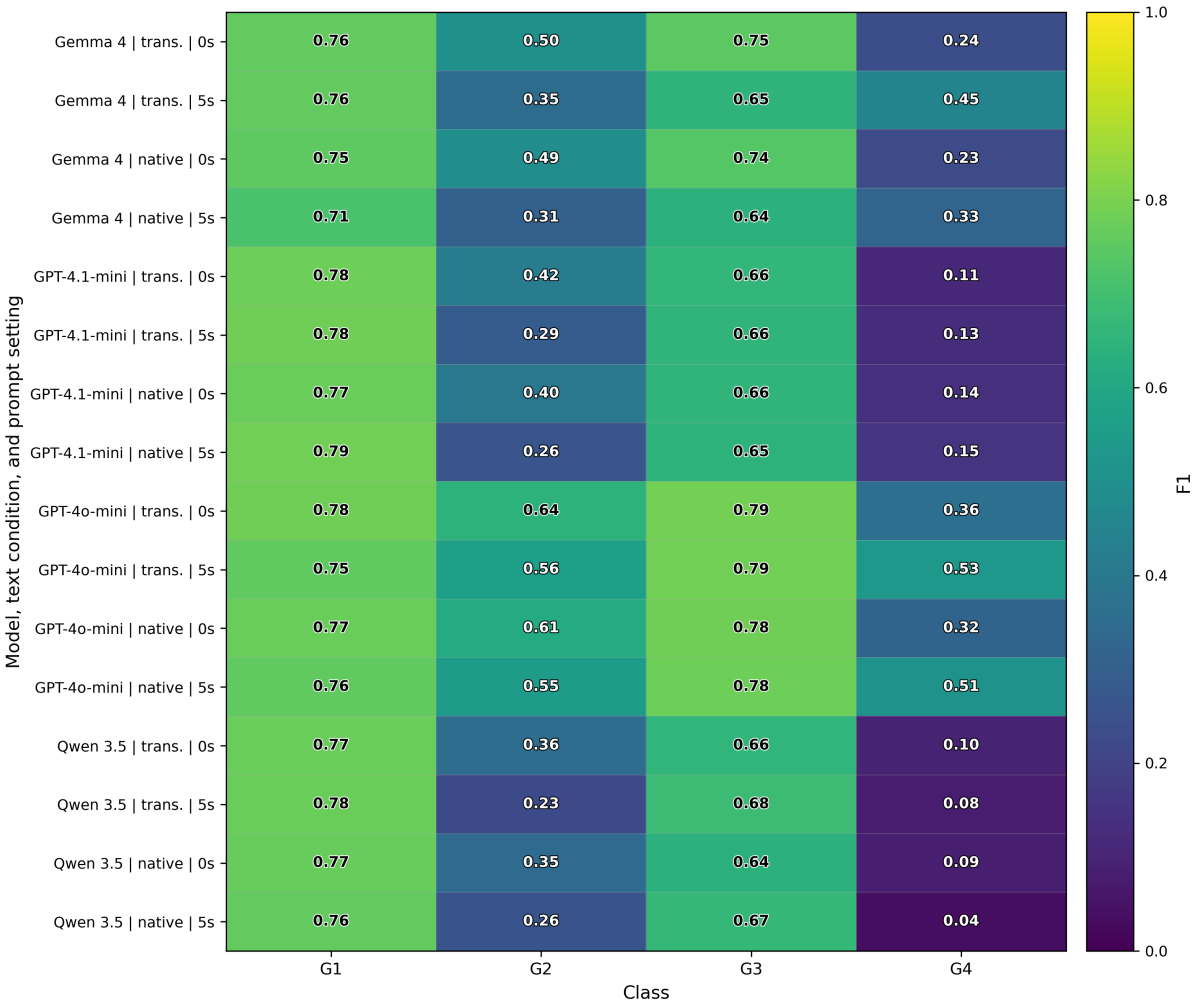


Figure 6: Classwise F1 for LLMs on grouped Task B.1. Values are computed on all 1,500 gold-standard posts. G1-G4 denote the grouped communication-function labels defined in Table 18.

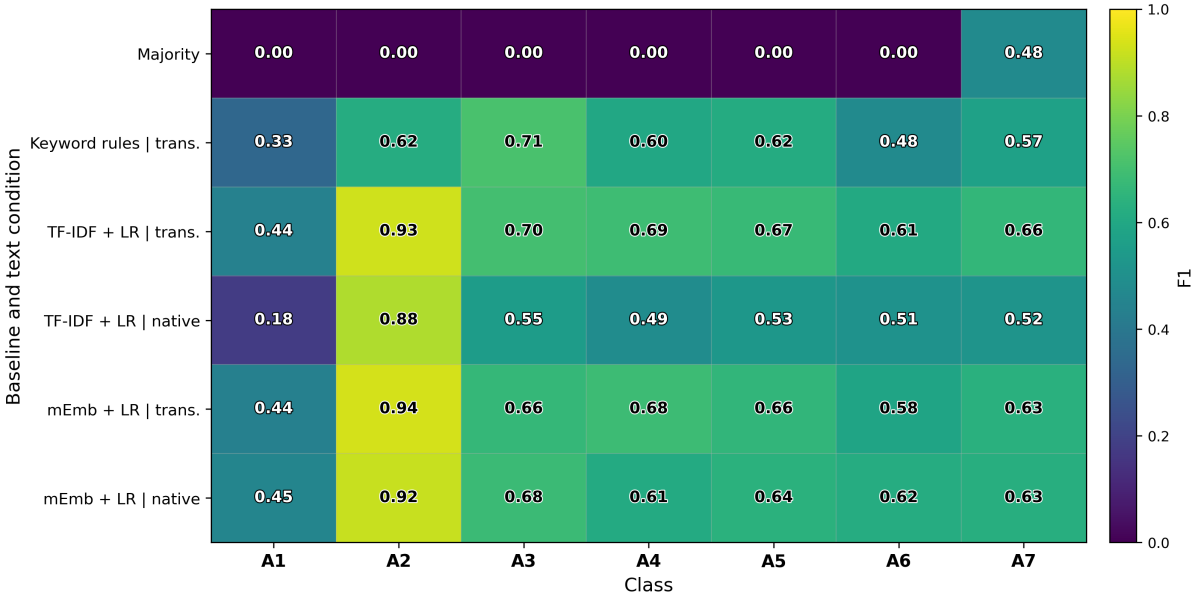


Figure 7: Classwise F1 for all baselines on Task A. Values are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. A1-A7 denote the Task A labels defined in Table 18.

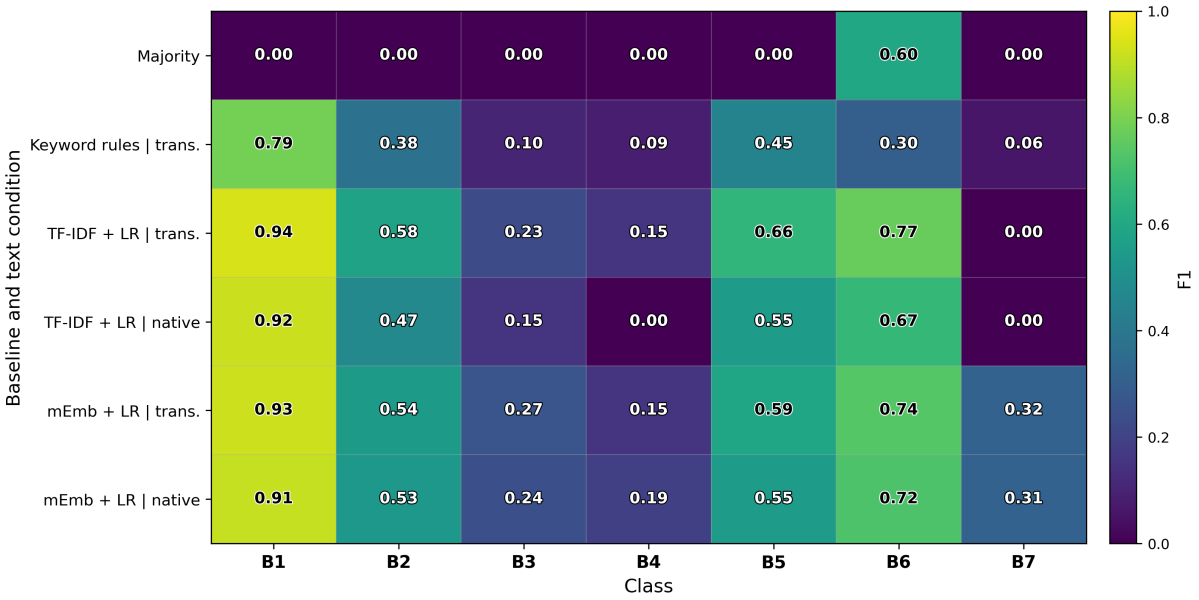


Figure 8: Classwise F1 for all baselines on Task B. Values are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. B1-B7 denote the Task B labels defined in Table 18.

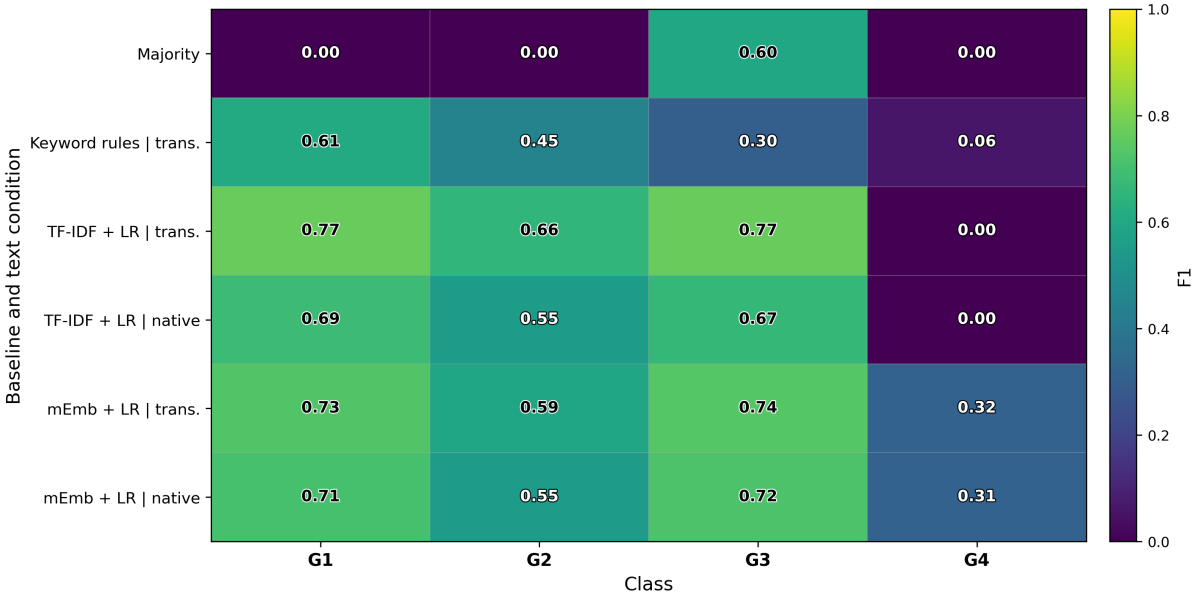


Figure 9: Classwise F1 for baselines on grouped Task B.1. Values are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. G1-G4 denote the grouped communication-function labels defined in Table 18.

Model	Text	Setting	Class	Precision	Recall	F1	Support
Gemma 4	translated	zero-shot	A1	0.481	0.585	0.528	65
Gemma 4	translated	zero-shot	A2	0.923	0.951	0.937	226
Gemma 4	translated	zero-shot	A3	0.738	0.764	0.751	140
Gemma 4	translated	zero-shot	A4	0.714	0.735	0.725	136
Gemma 4	translated	zero-shot	A5	0.469	0.842	0.603	209
Gemma 4	translated	zero-shot	A6	0.467	0.604	0.527	250
Gemma 4	translated	zero-shot	A7	0.888	0.384	0.536	474

Table 19: Classwise precision, recall, and F1 for the strongest LLM on Task A: Gemma 4, translated, zero-shot. Support is the number of gold examples in each class.

Model	Text	Setting	Class	Precision	Recall	F1	Support
GPT-4o-mini	translated	zero-shot	B1	0.706	0.059	0.108	205
GPT-4o-mini	translated	zero-shot	B2	0.442	0.530	0.482	164
GPT-4o-mini	translated	zero-shot	B3	0.069	0.667	0.124	21
GPT-4o-mini	translated	zero-shot	B4	0.200	0.273	0.231	11
GPT-4o-mini	translated	zero-shot	B5	0.671	0.621	0.645	446
GPT-4o-mini	translated	zero-shot	B6	0.791	0.788	0.790	638
GPT-4o-mini	translated	zero-shot	B7	0.333	0.400	0.364	15

Table 20: Classwise precision, recall, and F1 for the strongest LLM on Task B: GPT-4o-mini, translated, zero-shot. Support is the number of gold examples in each class.

Model	Text	Setting	Class	Precision	Recall	F1	Support
GPT-4o-mini	translated	five-shot	G1	0.875	0.663	0.755	401
GPT-4o-mini	translated	five-shot	G2	0.732	0.460	0.565	446
GPT-4o-mini	translated	five-shot	G3	0.673	0.950	0.788	638
GPT-4o-mini	translated	five-shot	G4	0.533	0.533	0.533	15

Table 21: Classwise precision, recall, and F1 for the strongest LLM on Task B.1: GPT-4o-mini, translated, five-shot. Support is the number of gold examples in each class.

Model	Text	Setting	Class	Precision	Recall	F1	Support
TF-IDF + LR	translated	Stratified 5-fold CV	A1	0.378	0.523	0.439	65
TF-IDF + LR	translated	Stratified 5-fold CV	A2	0.990	0.872	0.927	226
TF-IDF + LR	translated	Stratified 5-fold CV	A3	0.685	0.714	0.699	140
TF-IDF + LR	translated	Stratified 5-fold CV	A4	0.598	0.809	0.688	136
TF-IDF + LR	translated	Stratified 5-fold CV	A5	0.703	0.646	0.673	209
TF-IDF + LR	translated	Stratified 5-fold CV	A6	0.542	0.692	0.608	250
TF-IDF + LR	translated	Stratified 5-fold CV	A7	0.754	0.589	0.661	474

Table 22: Classwise precision, recall, and F1 for the strongest supervised baseline on Task A: TF-IDF + LR using translated text. Scores are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. Support is the number of gold examples in each class.

Model	Text	Setting	Class	Precision	Recall	F1	Support
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B1	0.944	0.907	0.925	205
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B2	0.436	0.726	0.545	164
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B3	0.177	0.524	0.265	21
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B4	0.103	0.273	0.150	11
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B5	0.647	0.543	0.590	446
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B6	0.816	0.669	0.736	638
Multilingual embeddings + LR	translated	Stratified 5-fold CV	B7	0.214	0.600	0.316	15

Table 23: Classwise precision, recall, and F1 for the strongest supervised baseline on Task B: multilingual embeddings + LR using translated text. Scores are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. Support is the number of gold examples in each class.

Model	Text	Setting	Class	Precision	Recall	F1	Support
Multilingual embeddings + LR	translated	Stratified 5-fold CV	G1	0.624	0.873	0.728	401
Multilingual embeddings + LR	translated	Stratified 5-fold CV	G2	0.647	0.543	0.590	446
Multilingual embeddings + LR	translated	Stratified 5-fold CV	G3	0.816	0.669	0.736	638
Multilingual embeddings + LR	translated	Stratified 5-fold CV	G4	0.214	0.600	0.316	15

Table 24: Classwise precision, recall, and F1 for the strongest supervised baseline on grouped Task B.1: multilingual embeddings + LR using translated text. Scores are computed from pooled out-of-fold predictions under stratified five-fold cross-validation. Support is the number of gold examples in each class.

Climate and Migration on German Political Party YouTube: A Corpus with Stance, Propaganda, and Divisiveness Annotation

Zaher Alkaei[1], Manfred Stede[1], Christoph M. Abels[2]
[1]Applied Computational Linguistics, University of Potsdam, Potsdam, Germany
[2]Department of Psychology, University of Potsdam, Potsdam, Germany
Correspondence: alkaei@uni-potsdam.de

Abstract

We present a German-language corpus for studying divisiveness in political social-media discourse: 7,277 parent–child comment pairs sampled from the YouTube channels of two ideologically opposed German parties, the AfD and Die Linke, on two contested policy domains, climate change and migration. Each pair is annotated on three dimensions: interaction quality (a five-class constructive–destructive $\times$ agreement–disagreement scale), stance (of parent and child toward the topic), and the presence of 14 propaganda techniques. We detail the full construction pipeline – channel selection, comment and transcript collection, pseudonymization, conversation-tree reconstruction, topic filtering, and stratified sampling – and a reproducible three-task annotation protocol: a 700-pair human gold standard validates an LLM-assisted pipeline that scales annotation to the full corpus. Descriptive analysis reveals pronounced party- and topic-specific patterns – most notably that the Migration $\times$ AfD stratum is the most hostile (54.5% of its pairs are Destructive Agreement) – while thread-level quality does not differ significantly between the two topics ($\chi^2(3) = 5.15$, $p = 0.16$). Only on climate are exchanges between opposing sides (a skeptic replying to a supporter) constructive on average; the same clash on migration remains destructive, and the most destructive exchanges overall occur between two opponents of the policy at issue.

1 Introduction

Online political discourse is increasingly polarized, yet modeling its dynamics remains methodologically challenging. Interlocutors negotiate disagreement within a pragmatic "disagreement space" (Jackson, 1992; Schumann and Oswald, 2024) where communication extends beyond propositional truth to the legitimacy, relevance, and appropriateness of speech acts. To capture both what interlocutors contest and how they contest it, we adopt the five-class interaction quality schema of Bassi et al. (2025), mapping comment-reply pairs along joint axes of agreement-disagreement and constructiveness-destructiveness.

A central motivation for this resource is that the surface act of disagreement does not by itself make an exchange destructive: opposing sides can still engage constructively. Consider a representative exchange from our corpus (translated from German):

Parent Comment

"Climate crisis, crisis, crisis… nag, nag, nag." *(stance: against climate action)*

Child Comment (Reply)

"Then why are crop failures from extreme weather getting more frequent, especially in Europe? The steady decline in birds, bees, and insects should be obvious to anyone …." *(stance: supports climate action)*

The child disagrees with the parent's dismissiveness, yet does so by pointing to concrete evidence rather than attacking. We label this pair *Constructive Disagreement* (interaction-quality score $+1.0$), with opposing parent and child stances and two techniques flagged on the child: *Appeal to Fear* (its warning of ecological consequences) and *Exaggeration/Minimisation* (the sweeping "obvious to anyone"). This is exactly the distinction that conventional toxicity-only or agreement-only annotation misses and that our three-layer scheme is designed to capture (see Table 5 for the full five-class spectrum). Conversely, agreement can be destructive: users who endorse one another while demeaning outsiders produce *Destructive Agreement*, which is in fact the dominant pattern in the corpus's most adversarial stratum, Migration $\times$ AfD (Section 5).

We build this resource from German-language YouTube comments on the channels of two ideologically opposed parties: the right-wing populist Alternative für Deutschland (AfD) and the left-wing Die Linke, focusing on two highly contested pol-

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 282–297
October 28, 2026 ©2026 Association for Computational Linguistics

icy domains, climate/energy (Coffé et al., 2026; Drecker, 2025) and migration/asylum. These channels concentrate exactly the kind of adversarial partisan exchange in which divisiveness is most salient, yet no German-language corpus annotates such discourse for fine-grained interaction quality.

The two parties anchor opposite poles of German partisan conflict. The AfD occupies the anti-migration extreme while left parties such as Die Linke hold pro-migration positions, structuring a polarized competition over immigration (Atzpodien, 2022); the same parties increasingly diverge on climate (Drecker, 2025), an issue that has grown deeply divisive in Germany (Zilles and Marg, 2023), with support for climate action falling fastest on the cultural right (Coffé et al., 2026). Social media, in turn, amplify the most hostile migration narratives well beyond their private support base (Levi et al., 2025). This makes these channels a fertile setting for studying how disagreement turns constructive or destructive.

This paper makes three primary contributions. First, we introduce a new German-language corpus of 7,277 YouTube comment pairs (2,000 threads) spanning two parties and two topics, labeled for interaction quality, stance, and propaganda techniques. Second, we describe a reproducible, LLM-assisted annotation methodology validated against a 700-pair human gold standard, and make the full annotation guidelines and prompts available. Third, we provide a descriptive analysis characterizing how constructive and destructive dynamics, stance, and rhetorical techniques distribute across party and topic channels.

2 Related Work

2.1 Interaction Quality in Polarized Platform Spaces

Computational approaches to online controversy have transitioned from macro-level network segregation metrics (Conover et al., 2011; Garimella et al., 2018) to fine-grained conversational processing. Micro-level studies analyze textual affect to show that controversial comments express more anger than non-controversial ones (Chen et al., 2023), or quantify controversy intensity by combining stance antagonism with sentiment, offensiveness, and sarcasm (Wang et al., 2023). Grounded in pragma-dialectical theory, we take the quality of disagreement to turn on the balance of hedging and certainty (De Kock and Vlachos, 2021), resistance to toxic escalation (Seckin et al., 2026), and argument and deliberative quality (Falk, 2025).

These conversational dynamics are heavily shaped by platform architecture. Asynchronous, text-based spaces like YouTube systematically exhibit lower mutual understanding and higher perceived conflict than spoken discourse (Bevis et al., 2026), and their architecture drives radicalization dynamics through user migration toward more extreme channels (Ribeiro et al., 2020). In the German far right, the comment sections of AfD, Pegida, and Identitarian channels increasingly converge on shared user communities (Rauchfleisch and Kaiser, 2020). Within such preference-driven, like-minded spaces, partisan forums exhibit lower toxic outrage than issue-driven mainstream arenas (Jakob et al., 2023). To capture these dynamics, we operationalize "divisiveness" as the destructive pole of the agreement and constructiveness axes in Bassi et al. (2025)'s schema.

2.2 Stance and Rhetorical Technique

We annotate stance and propaganda techniques as complementary layers to interaction quality rather than in isolation, since both are theoretically tied to divisiveness. Stance – target-specific favorability (Mohammad et al., 2016) – is often orthogonal to conversational style: ideological opponents can engage in constructive debate, while aligned users frequently reinforce one another's hostility. We therefore annotate stance (Against, Neutral, Support) at both the parent and child level, capturing configurations such as constructive exchanges between opposing stances (Bassi et al., 2025).

Rhetorical manipulation is likewise tightly coupled with destructive interaction. While explicit toxicity is easily flagged by lexical cues, covert ideological appeals and democratic norm violations require finer rhetorical distinctions (Stoll et al., 2020). Grounded in political communication theory (Zompetti, 2015), we adopt the 14-technique propaganda taxonomy of Bassi et al. (2025) and apply it to German partisan discourse, annotating the divisive rhetorical devices present in each child comment.

2.3 LLM-in-the-Loop Annotation

LLMs have emerged as scalable annotators for political text, matching human performance on tasks from stance and frame detection (Gilardi et al., 2023) to propaganda labeling (Jose et al., 2026). However, relying on LLMs introduces systematic biases and domain-generalization limits, particu-

larly for lower-resource languages (Jadhav et al., 2025). This raises the question of when LLM annotations can responsibly substitute for human ones (Calderon et al., 2025). We therefore adopt a hybrid strategy: we quantify LLM-to-human agreement on a manually annotated 700-pair gold standard before using `claude-opus-4-6` to scale annotations to the remaining 6,577 pairs.

2.4 Related Corpora

Prior corpora typically target one dimension at a time: constructiveness on Wikipedia talk pages (De Kock and Vlachos, 2021) or news comments (Kolhatkar et al., 2020), multilingual stance in voting-advice applications (Vamvas and Sennrich, 2020), or offensive language in German social media (Wiegand et al., 2018). Multi-label resources exist – GermEval 2021 scores toxicity, engagement, and fact-claiming (Risch et al., 2021), and Wang et al. (2023) fuse stance, sentiment, offensiveness, and sarcasm into one controversy score for Weibo – but neither labels the quality of an individual exchange. The closest corpus is that of Bassi et al. (2025), whose three-task schema we adopt; theirs covers English-language US immigration discourse, a scope its authors note is confined to one language and topic. Ours annotates German comment pairs across two policy domains and two parties' channels.

3 Dataset

3.1 Channel Selection and Collection

We target two German parties at opposite ends of the parliamentary spectrum, the right-wing populist *Alternative für Deutschland* (AfD) and the left-wing *Die Linke*, and collect from each party's official channel and its parliamentary-group (*Fraktion*) channel: `@AfDTV` and `@AfDFraktionimBundestag` for the AfD, and `@DIELINKE` and `@dielinkebt` for Die Linke. Using the YouTube Data API v3, we retrieved for every uploaded video its metadata, all comments, and, where available, the German transcript, over 06–13 March 2026. This yielded 14,280 videos (10,856 AfD; 3,424 Die Linke) and 4,638,722 comments; transcripts were available for 12,621 videos. A per-channel breakdown of the topic-relevant subset is given in Appendix A.1.

3.2 Pseudonymization

All author names were collected, sorted, and replaced with sequential pseudonyms (`User_00001`, `User_00002`, ...); the same deterministic mapping was applied to `@`-mentions inside comment text, and erroneous double-`@@` patterns introduced by YouTube's formatting were normalized. Names were mapped separately for each party's channels, so an author is identified by the (party, pseudonym) pair: 476,330 identities were pseudonymized, 386,213 on the AfD channels and 90,117 on the Die Linke channels. All subsequent processing and the public release use only the pseudonymized text.

3.3 Conversation Reconstruction

The YouTube API returns every reply as a direct child of the thread root, discarding the true addressee when users reply to one another deep in a thread. Following Bassi et al. (2024), we reconstruct the reply structure: processing each video's comments chronologically, we maintain a global registry mapping every author to their most recent comment, and for each reply we take its last `@`-mention as the intended addressee, re-attaching the reply to that author's latest comment (falling back to the raw parent when the mention cannot be resolved). Each comment thereby receives a resolved parent and a conversation depth. From every resolved link we form a parent–child pair, discarding *mismatched-target* pairs, in which the child's `@`-mentions refer only to third parties, so that every retained pair is an unambiguous dyadic exchange. Bassi et al. (2025) manually validated this reconstruction procedure on English YouTube data, finding only 30 incorrect links among 2,387 child comments (98.7% accuracy).

3.4 Thematic Filtering via Topic Modeling

To isolate climate and migration discussions, we fit a BERTopic model (Grootendorst, 2022) on the 12,621 available transcripts, using `paraphrase-multilingual-MiniLM-L12-v2` embeddings and class-based TF-IDF weighting, yielding 52 topics plus an outlier class. We manually mapped topics to two domains by their top keywords (Table 1); each video inherits its dominant topic, and videos assigned to neither domain were discarded.

A blind manual check of 75 videos found 75.0% of assigned videos substantially on-topic, with errors drawn mainly from adjacent policy domains; the off-topic remainder inflates the neutral-stance mass but contributed no opposing-stance pairs in that sample (Appendix A.2).

Table 1: Thematic categories and representative BERTopic keywords (translated from German).

Category	Topic IDs	Representative Keywords
Migration	1, 9, 11, 33, 35	*migrants, refugees, asylum, borders*
Climate	3, 4, 36	*CO2, climate, energy transition, nuclear power*

3.5 Pair Extraction and Stratified Sampling

After topic filtering and pair extraction, the topic-relevant corpus comprises 1,143 videos and 385,320 threads (full per-channel statistics in Appendix A.1). Because the large majority of threads are single root comments with no replies, pair extraction yields 125,967 clean parent–child pairs, concentrated in just 39,861 threads (≈10% of all threads), which form the pool eligible for sampling. This pool is highly imbalanced: 91.7% of eligible threads come from AfD channels, and Climate × Die Linke contributes only 723 (1.8%). Annotating it in full would therefore yield a corpus dominated by a single stratum, confounding party and topic effects with channel volume, while the smallest stratum caps any balanced design at 723 threads per cell. To keep the party × topic contrasts that motivate this resource directly comparable, we instead drew 2,000 threads with a fixed random seed, 500 per party–topic stratum, from this eligible pool (Table 2). This produced 7,299 pairs; removing 22 duplicate pairs (all in the Climate × Linke stratum, arising from comments shared across reconstructed threads) leaves a final corpus of 7,277 unique, clean pairs.

Table 2: Stratified sampling design and yield of clean, unique pairs.

Stratum	Threads	Comments	Clean Pairs
Climate × AfD	500	2,632	1,999
Climate × Linke	500	2,204	1,420
Migration × AfD	500	2,691	1,984
Migration × Linke	500	2,683	1,874
Total	2,000	10,210	7,277

4 Annotation

4.1 Annotation Schema

Following Bassi et al. (2025), every comment pair is annotated on three tasks. **Interaction quality** is a five-class ordinal label on the child's contribution relative to its parent, jointly encoding communication style (constructive vs. destructive) and relation (agreement vs. disagreement): Constructive Disagreement (+1.0), Constructive Agreement (+0.5), Neutral/Rephrase (0.0), Destructive Agreement (−0.5), and Destructive Disagreement (−1.0). **Stance** toward the video's topic (Against, Neutral, Support) is annotated independently for the parent and the child comment. **Propaganda** is a 14-label multi-label set of divisive rhetorical techniques (Zompetti, 2015; Bassi et al., 2025), applied to the child comment. Table 3 gives the operational definitions of the five classes.

Table 3: Interaction Quality category definitions and score mappings.

Category	Score	Definition and Linguistic Criteria
Constructive Disagreement	+1.0	Expresses disagreement while preserving conditions for mutual understanding; evaluates specific claims with reasons or examples.
Constructive Agreement	+0.5	Supports the parent while offering substantive reasons, data, or arguments that advance the dialogue.
Neutral/Rephrase	0.0	Neutral entries serving conversational flow without taking sides, including information requests or topic-shifting.
Destructive Agreement	−0.5	Agrees with the parent but demeans/invalidates opposing viewpoints to build exclusive in-group solidarity.
Destructive Disagreement	−1.0	Obstructs dialogue using hostile tone, personal attacks, or aggressive language; dehumanization triggers this score.

4.2 Annotation Guidelines

The guidelines adapt those of Bassi et al. (2025) to the climate and migration domains, adding supplementary decision rules for both the human and the LLM annotators. Interaction quality is judged along two axes: functional alignment (agreement, disagreement, neutral) and communication style (constructive, destructive). Stance is assessed against topic-specific policy targets, independently for parent and child. Propaganda flags the presence of 14 rhetorical devices in the child comment, with disambiguation rules for closely related techniques (e.g. Exaggeration/Minimisation vs. Loaded Language, or Repetition vs. parallel sentence structure). A condensed version of the guidelines and system prompts is provided in Appendix A.3; the complete prompts are released with the corpus.

4.3 Human Gold Standard

Two annotators independently labeled a stratified sample of 700 pairs (175 per stratum) through a custom local tool that presented each pair with full context (party, topic, thread, comment levels, and timestamps) and captured all three tasks simultaneously; pairs in which either comment was shorter than 20 characters, or which were flagged as mismatched-target, were filtered out beforehand. The 75 pairs (10.7%) on which the annotators agreed exactly across all three tasks entered the gold standard directly; the remaining 625 (89.3%) were adjudicated to a single label per task through discussion. Table 4 reports the resulting label distribution.

Table 4: Gold-standard label distribution ($n = 700$).

Task / Label	Gold Count	Gold %
Interaction Quality (IQ)		
Constructive Disagreement (+1.0)	155	22.1%
Constructive Agreement (+0.5)	69	9.9%
Neutral/Rephrase (0.0)	119	17.0%
Destructive Agreement (−0.5)	160	22.9%
Destructive Disagreement (−1.0)	197	28.1%
Child Stance		
Support	54	7.7%
Neutral	568	81.1%
Against	78	11.1%
Propaganda (14 classes)		
Top: *Loaded Language*	332	47.4%

Prior to adjudication, exact agreement on the 5-class ordinal IQ task reached 56.9% (unweighted Cohen's $\kappa = 0.456$, quadratic-weighted $\kappa = 0.494$). Parent and child stance agreement reached 87.0% ($\kappa = 0.666$) and 87.4% ($\kappa = 0.640$). Multi-label propaganda detection yielded a mean Jaccard index of 0.335 (mean multi-label $F_1 = 0.392$). The interaction-quality agreement exceeds the $\kappa = 0.37$ that Bassi et al. (2025) report for the same schema on English YouTube comments; both figures reflect the difficulty of operationalizing constructive versus destructive communication in informal online discourse. See Appendix A.6 for expanded agreement matrices.

4.4 Representative Examples

To illustrate the 5-class interaction quality spectrum, Table 5 provides representative gold-standard comment pairs alongside their operationalized stance and rhetorical markers.

4.5 LLM-Assisted Annotation

To scale annotation beyond the gold standard, we label the remaining pairs with `claude-opus-4-6` via the Anthropic API, using four independent task-specific calls per pair: a structured tool-use call for interaction quality, separate system-prompted calls for child and parent stance, and a free-text call for propaganda techniques validated against the 14-label vocabulary. Before scaling, we validated the pipeline on the 700-pair gold standard, measuring the LLM's agreement with the human labels across all three tasks (Table 6).

The LLM achieved moderate agreement on interaction quality (quadratic-weighted $\kappa = 0.560$; unweighted 0.550) and fair-to-moderate agreement on both stance layers: child stance ($\kappa = 0.428$) and parent stance ($\kappa = 0.435$). For child stance, the prevalent *Neutral* class was substantially easier to classify ($F_1 = 0.791$) than the polarized *Against* ($F_1 = 0.515$) or *Support* ($F_1 = 0.537$) postures. Re-annotating a random 200-pair sample under identical settings reproduced 93.0% of interaction-quality labels (quadratic-weighted $\kappa = 0.906$) and 96.0% of child stance labels ($\kappa = 0.930$), so the moderate agreement with the gold standard reflects the subjectivity of the task rather than instability in the model.

Following validation, the pipeline was run on the remaining pairs, yielding a final corpus of 7,277 pairs (700 human gold standard, 6,577 LLM-annotated).

5 Thread-Level Score Aggregation and Corpus Analysis

We aggregated pair-level interaction quality scores to the thread level by calculating the mean `InteractionScore` for each (`VideoID`, `ThreadID`) cluster. Following Bassi et al. (2025), average thread scores were mapped into four quality bins (Table 7).

Breaking these thread-quality bins down by stratum (Table 8) exposes a pattern the aggregate hides: the AfD strata concentrate in the *Moderately Destructive* bin (54–60%) with comparatively few *Highly Destructive* threads, whereas the Die Linke strata are more polarized, carrying both more Constructive *and* more Highly Destructive threads. The mean thread scores nonetheless confirm that the AfD strata are the more destructive overall. Bin names and thresholds follow Bassi et al. (2025), who define them on a divisiveness scale, so they

Table 5: Representative comment-reply pairs illustrating the five-class interaction quality (IQ) schema (translated from German). Topic: M = migration, C = climate.

Topic	IQ	Stance	Parent Comment	Child Comment	Rhetorical Markers & Notes
M	-1.0	P: Supp., C: Agst.	"I'm with the SPD and would take in refugees. Unfortunately rejected due to lack of space."	"@User_10960 That's not a problem – you could just move out and sleep on the street..."	**Destructive Disagreement**. Uses sarcasm to mock the parent's welcome stance. (*Loaded Language*).
M	-0.5	P: Agst., C: Agst.	"Bed, bread, and soap. Then they'll leave on their own, and we won't have to spend billions..."	"I'd be in favor of that too, but I'm worried they'd become even more criminal..."	**Destructive Agreement**. Affirms parent but escalates threat with unevidenced outgroup stereotypes. (*Appeal to Fear*, *Causal Oversimplification*, *Loaded Language*, *Name-Calling*).
M	0.0	P: Neut., C: Neut.	"Thanks for not giving in to the rightward shift, unlike the Greens."	"Please explain that."	**Neutral/Rephrase**. Requests clarification without taking a clear side or shifting tone. (*Doubt*).
C	$+0.5$	P: Supp., C: Supp.	"The 1.5°C target can no longer be held? 1.5 has been dead for a long time already."	"And no longer reversible! 19 degrees in November, yet another sad new record."	**Constructive Agreement**. Validates parent and introduces local weather anomaly data. (*Appeal to Fear*, *Loaded Language*).
C	$+1.0$	P: Agst., C: Supp.	"What qualifications does she actually have to make this claim? Is this backed up by data?"	"Yes, it's in the IPCC report and many other studies on this..."	**Constructive Disagreement**. Challenges parent directly by referencing institutional consensus. (*Appeal to Authority*).

Table 6: LLM alignment with the human gold standard ($n = 700$). IQ: interaction quality; Prop.: propaganda.

Task	Metric	Score
IQ	Quad-weighted κ / Unweighted κ	0.560 / 0.550
	Mean Absolute Error (MAE)	0.366
	Pearson correlation (r)	0.560
Stance (child)	Unweighted κ / Macro-F_1	0.428 / 0.615
	Class F_1 (Against / Neutral / Support)	0.515 / 0.791 / 0.537
Stance (parent)	Unweighted κ / Macro-F_1	0.435 / 0.645
	Class F_1 (Against / Neutral / Support)	0.544 / 0.761 / 0.631
Prop.	Macro-F_1 / Hamming Loss	0.460 / 0.118

index the composite score rather than destructiveness alone. Because Destructive Agreement carries -0.5, a thread of purely destructive agreement cannot fall below that value: the Destructive-Agreement-dominated AfD strata therefore concentrate in the *Moderately Destructive* bin, while the Die Linke strata, richer in Destructive Disagreement, reach *Highly Destructive* more often despite fewer destructive pairs overall.

Table 7: Thread-level quality bins and corpus distribution ($n = 2,000$ threads).

Mean Score	Quality Bin	Threads	%
$\geq +0.25$	Constructive	412	20.6%
$[-0.25, +0.25)$	Slight / Neutral	487	24.4%
$[-0.75, -0.25)$	Mod. Destructive	829	41.5%
< -0.75	Highly Destructive	272	13.6%

5.1 Stratum and Rhetorical Dynamics

By tracking interaction quality and stance across our stratified design, we observed stark contrasts between political channels (Table 9). The same distribution is visualized in Appendix Figure 1.

A structural divergence emerged along the agreement axis: discussions on AfD channels are dom-

Table 8: Thread-level quality-bin composition per stratum (% of the 500 threads in each stratum) and mean thread interaction score. The mean ranges from -1.0 (every reply Destructive Disagreement, maximally divisive) through 0 (neutral) to $+1.0$ (every reply Constructive Disagreement, maximally constructive).

Stratum	Con.	Sl./N.	Mod.D.	Hi.D.	Mean
Climate × AfD	12.8	23.6	54.0	9.6	−0.283
Climate × Linke	31.6	26.6	26.0	15.8	−0.077
Migration × AfD	10.6	20.0	59.8	9.6	−0.325
Migration × Linke	27.4	27.2	26.0	19.4	−0.129

inated by agreement (50.3% for climate, 58.0% for migration), while comments on Die Linke are primarily disagreeing (54.2% for climate, 62.5% for migration). This suggests that AfD comment sections are spaces where users predominantly reinforce one another's positions, whereas Die Linke channels host more contestation.

The two collapses (Table 10) show that party shapes both axes, and that agreement does not imply civility. On the *relation* axis, AfD channels skew toward agreement and Die Linke toward disagreement irrespective of topic; on the *communication style* axis, destructiveness peaks in the AfD strata (72.1% for Migration × AfD) while the Die Linke strata are the most constructive (38.2% for Climate × Linke). Migration × AfD combines the highest agreement (58.0%) with the highest destructiveness, driven by Destructive Agreement – 54.5% of its pairs – in which users reinforce hostile shared positions rather than engage opposing ones. This tone is mirrored in the prevalence of specific propaganda techniques (Table 11).

Table 11 breaks technique prevalence down by stratum: emotive devices dominate everywhere (Loaded Language appears in roughly half of all pairs), while fear appeals and flag-waving concentrate in Migration × AfD (17.4% and 11.1%), the stratum that also shows the highest Destructive Agreement.

Child-stance distributions mirror the party divide (Table 9). Neutral is the most common label in three of the four strata, from 51.9% in Migration × Linke to 72.8% in Climate × Linke, reflecting that many comments beneath a topic-labeled video take no explicit policy position of their own. The exception is Migration × AfD, where Against-stance is by far most common: 56.8% of children oppose immigration (Neutral falls to 37.6%), whereas Support concentrates in the Die Linke strata (16.5% for climate and 22.1% for migration, versus 7.1% and 5.7% on the AfD channels). Stance therefore tracks channel ideology closely, while the large neutral mass underscores that topic membership is defined at the video rather than the comment level.

Climate: cross-cutting dialogue versus same-stance reinforcement. Grouping pairs by the *stance relation* between parent and child reveals a pattern specific to climate (Table 12). Only on climate are *opposing*-stance exchanges – a skeptic engaging a supporter – constructive on average (mean $+0.08$, 53% constructive; $n = 340$), the single positive relation in either topic; the analogous migration clash stays destructive (mean -0.07), and the two differ significantly (Welch $t = 2.28$, $p = 0.02$). This holds within both channels: opposing-stance climate exchanges are constructive on average on the AfD ($+0.04$) and Die Linke ($+0.12$) channels alike (Appendix A.5). Divisiveness instead concentrates in *same-stance* exchanges, but unevenly across them (Table 12). The destructive engine is *hostile agreement between opponents*: pairs in which both sides are Against – skeptic replying to skeptic on climate, anti-migration to anti-migration – are the single most destructive configuration we report (mean -0.39 on climate, -0.41 on migration; roughly 88% destructive). Neutral-to-neutral pairs are only mildly destructive (-0.17 / -0.11) but by far the most frequent, so their volume masks how hostile the same-stance relation otherwise is: excluding them, the same-stance mean falls to -0.29 on climate and -0.38 on migration. Support-to-support pairs, a sub-case of same stance rather than a relation in its own right, are by contrast mildly *constructive* ($+0.13$ / $+0.03$) though rare: it is opponents reinforcing hostility, not allies agreeing, that produces divisiveness. This destructiveness is concentrated on the AfD side (same-stance mean -0.31 vs. -0.03 for climate), and not merely because AfD channels host more Against–Against exchanges: both Against–Against and Neutral–Neutral pairs are harsher there, the latter averaging -0.28 against -0.05 on Die Linke. Climate is thus the domain where cross-cutting dialogue still functions; on migration, the same skeptic–supporter encounter turns destructive. This tracks the two topics' rhetorical signatures: climate discourse leans on *Doubt* (13.6% of pairs vs. 9.7% for migration) more than *Appeal to Fear* (9.6% vs. 13.8%), and climate skepticism is concentrated on the AfD chan-

Table 9: Pair-level interaction quality and child stance distributions across strata (n = 7,277 unique pairs). IQ labels: CD = Constructive Disagreement ($+1.0$), CA = Constructive Agreement ($+0.5$), N = Neutral (0.0), DA = Destructive Agreement (-0.5), DD = Destructive Disagreement (-1.0).

		Interaction Quality (5-Class)					Child Stance		
Stratum	n	CD	CA	N	DA	DD	Against	Neutral	Support
Climate × AfD	1,999	301	89	257	916	436	596	1,262	141
Climate × Linke	1,420	387	156	244	251	382	152	1,034	234
Migration × AfD	1,984	244	69	241	1,081	349	1,126	745	113
Migration × Linke	1,874	552	139	249	315	619	487	973	414

Table 10: Two aggregations of the five-class interaction labels, by stratum (row % of pairs). *Relation*: agreement vs. disagreement; *Style* (communication style): constructive vs. destructive. Neut. (Neutral/Rephrase) is shared by both aggregations. Strata: C = Climate, M = Migration; A = AfD, L = Die Linke.

	Relation		Style		
Stratum	Agree	Disag.	Constr.	Destr.	Neut.
C × A	50.3	36.9	19.5	67.6	12.9
C × L	28.7	54.2	38.2	44.6	17.2
M × A	58.0	29.9	15.8	72.1	12.1
M × L	24.2	62.5	36.9	49.8	13.3

nels, where 29.8% of comments oppose climate action versus 10.7% on Die Linke.

5.2 Temporal and Thematic Trends

Analyzing thread quality by topic from 2019 onward (Table 13), climate threads carry a higher constructive share than migration threads in most years, though migration briefly overtakes climate in 2022–2023; both topics drop sharply in 2023 and recover by 2024, after which the gap narrows to within three points by 2026. A full breakdown by party × topic stratum and year is given in Appendix A.4 (Table 19). Despite these fluctuations, the overall thread-quality distributions do not differ significantly between the two topics ($\chi^2(3) = 5.15$, $p = 0.16$). Because the sample is not balanced by party within years (it is balanced by party overall), these year-on-year movements describe the sampled threads rather than a development in the discourse itself (Section 8).

6 Corpus Format and Availability

The corpus is released as JSON Lines, one record per parent–child pair. Each record carries the pair identifier; the pseudonymized parent and child text with their author identifiers, conversation levels, and timestamps; the source video and thread identifiers; the party and topic stratum; and the three annotation layers – the interaction-quality score and label, parent and child stance, and the set of propaganda techniques on the child comment. The human gold standard and the LLM-annotated remainder are released as separate files, so that the 700-pair gold subset can be used on its own. The full dataset, annotation guidelines, system prompts, and the construction and annotation code are publicly released.[1]

7 Conclusion

We presented a new German-language resource for studying divisiveness in political social-media discourse: a corpus of 7,277 parent–child comment pairs sampled from AfD and Die Linke YouTube channels on climate and migration, annotated for interaction quality, stance, and propaganda techniques. We described the full construction pipeline – channel selection, comment and transcript collection, pseudonymization, conversation reconstruction, topic filtering, and stratified sampling – and a reproducible three-task annotation protocol validated against a 700-pair human gold standard. Descriptive analysis reveals pronounced party- and topic-specific patterns, most notably the predominance of Destructive Agreement in the Migration × AfD stratum. Zooming in on climate, we find that cross-cutting skeptic–supporter exchanges are the only stance relation that stays constructive on average, on both parties' channels, while same-stance exchanges carry most of the divisiveness.

8 Limitations

First, 90.4% of the corpus is annotated with `claude-opus-4-6` rather than by humans and its agreement with the human gold standard is moderate ($\kappa = 0.550$). The human–human baseline is itself moderate ($\kappa = 0.456$) and exceeds the

[1] https://github.com/zaheralkaei/climate_migration_german_political_yt

Table 11: Propaganda technique prevalence per stratum (% of pairs); the highest value in each row is shown in bold.

Technique	Climate × AfD	Climate × Linke	Migration × AfD	Migration × Linke
Appeal_to_Authority	3.2%	**4.7%**	2.3%	2.6%
Appeal_to_Time	4.4%	2.9%	**6.1%**	2.7%
Appeal_to_fear-prejudice	11.8%	6.5%	**17.4%**	10.0%
Bandwagon,Reductio_ad_hitlerum	1.8%	2.5%	2.1%	**2.9%**
Black-and-White_Fallacy	3.7%	4.7%	5.0%	**5.4%**
Causal_Oversimplification	9.5%	13.0%	11.3%	**13.8%**
Doubt	**13.6%**	13.5%	8.6%	10.8%
Exaggeration,Minimisation	25.2%	25.0%	26.3%	**27.6%**
Flag-Waving	7.7%	3.9%	**11.1%**	5.0%
Loaded_Language	51.9%	44.2%	**53.6%**	48.2%
Name_Calling,Labeling	**38.8%**	30.0%	34.2%	35.3%
Repetition	3.3%	**3.9%**	3.3%	2.2%
Slogans/Thought-terminating_Cliches	9.6%	10.0%	**10.2%**	9.1%
Whataboutism,Straw_Men	2.7%	2.3%	2.1%	**5.0%**

Table 12: Interaction quality by parent–child stance relation, aggregated across both parties, with the same-stance relation decomposed by shared stance. Mean interaction score (range -1 to $+1$) and pair count n. Opposing (skeptic vs. supporter) is the only relation constructive on average, and only on climate; hostile Against–Against agreement is the most destructive of the configurations reported here.

	Climate		Migration	
Stance relation	Mean	n	Mean	n
Opposing	$+0.08$	340	-0.07	583
Same stance	-0.19	2,276	-0.24	2,377
Against–Against	-0.39	401	-0.41	1,028
Neutral–Neutral	-0.17	1,780	-0.11	1,263
Support–Support	$+0.13$	95	$+0.03$	86
One side neutral	-0.22	803	-0.18	898

Table 13: Constructive-thread share (% of threads in the Constructive bin) by year and topic (year of video publication), 2019–2026; thread counts n in parentheses. Years before 2019 contain too few threads (< 15) to be informative.

Year	Climate (n)	Migration (n)
2019	28.9 (45)	7.0 (43)
2020	11.8 (68)	6.8 (73)
2021	23.1 (104)	12.8 (47)
2022	11.9 (151)	14.3 (56)
2023	7.0 (129)	8.9 (123)
2024	34.2 (73)	19.3 (228)
2025	28.5 (361)	27.0 (352)
2026	25.0 (60)	24.6 (65)

$\kappa = 0.37$ reported for the same schema by Bassi et al. (2025); the model is highly self-consistent under re-annotation (93.0% exact on interaction quality, quadratic-weighted $\kappa = 0.906$), so little of the residual gap is model noise; and the two annotators differ mainly by a uniform strictness offset rather than a content-linked bias (Appendix A.6). The remaining disagreement therefore reflects the inherent subjectivity of interaction-quality annotation rather than instability in the pipeline. Fine-grained comparisons between individual classes should be read with that subjectivity in mind.

Second, topic filtering operates at the video level, and the blind check in Section 3.4 shows that even video-level assignment is imperfect; individual comments may drift further still. This labeling is therefore a rough filter for locating discussions related to climate and migration, not a precise topic label for each individual comment.

Third, the stratified sampling design imposes an artificial balance across parties and topics that does not reflect the natural class distribution of the source channels; we release the full eligible-thread pool so the pipeline can be scaled to the unbalanced corpus where representativeness matters more than cross-stratum comparability.

Fourth, aggregating the ordinal interaction-quality classes to the thread level requires numeric values. Following Bassi et al. (2025) we use $(+1.0, +0.5, 0, -0.5, -1.0)$, the equal-interval coding of the class ranks rescaled to $[-1, +1]$, so the larger magnitude of the disagreement classes follows from the class ordering rather than a separate weighting choice; the bin thresholds are likewise inherited. Our central result is robust to this: recoding by communication style alone (constructive $= +1$, neutral $= 0$, destructive $= -1$), which discards the agreement axis, still leaves opposing-stance climate exchanges the only relation with a positive mean. We release the categorical labels alongside the scores, so alternative aggregations can be derived.

Fifth, the sample is balanced across party and

topic overall but not within years, so its composition shifts over the period: pairs from 2020 are 99% AfD, whereas pairs from 2025 are 83% Die Linke. Corpus-wide year-on-year movements in destructiveness, agreement, and technique use largely track this shifting composition rather than a change in discourse. Some within-party series do move – the constructive share of Migration × AfD threads rises from 4.9% in 2019 to 21.0% in 2025 – but within a party the sampled videos also change from year to year, so the temporal breakdowns in Table 13 and Appendix A.4 are descriptive summaries of the sample rather than evidence of development over time; establishing genuine trends would require a temporally stratified sample.

9 Ethical Considerations

Data and privacy. The corpus is built exclusively from publicly posted YouTube comments, collected through the official Data API v3 in accordance with its terms of service. Prior to any annotation or release, all author names were replaced with sequential pseudonyms and @-mentions rewritten with the same deterministic mapping, so that no usernames or account handles remain in the released text; we make no attempt to re-identify authors and release no author profile metadata.

Intended use and risks. The corpus contains, in parts, hostile, offensive, and propagandistic German-language political speech, and is released for research on online discourse, content moderation, and computational social science. It is not intended as training data for generating such content, and the propaganda-technique labels characterize rhetorical form, not the truth or endorsement of any claim.

Annotation. The gold standard was produced by two trained annotators familiar with German political discourse; no crowd workers were employed. Annotators saw each pair together with its thread context, including the source channel, because resolving sarcasm, deictic reference, and the target of a stance claim requires that context. Annotators were therefore not blind to party, which carries a risk of confirmation bias, with comments from a disfavored party read as more destructive by default. Three features of the design mitigate this. The guidelines instruct annotators to judge communication style rather than ideological alignment (Appendix A.3.1); 89.3% of pairs were reconciled through joint adjudication, which surfaces systematic individual tendencies; and the LLM annotations that cover the remaining 90.4% of the corpus are party-blind, since each prompt receives only the parent and child comment text, with no channel, party, or video metadata. Party information is thus visible only for the 9.6% human-annotated subset. To test for a party-linked annotation effect directly, we compared the human gold labels against the party-blind LLM labels on the same pairs. Had annotators systematically read AfD comments as more destructive, the human scores would fall below the LLM scores on the AfD strata specifically. They do not: the mean human-minus-LLM interaction score is $+0.013$ on the AfD strata and -0.050 on the Die Linke strata (difference 0.063 on a $[-1, +1]$ scale; Welch $t = 1.19$, $p = 0.24$), so the small observed difference runs opposite to the hypothesized direction and is not statistically significant. The party-blind LLM in fact labels a larger share of AfD pairs destructive (67.4%) than the human annotators do (58.3%). This is a consistency check rather than proof of absence, since the LLM is not a neutral reference and party affiliation is sometimes inferable from the comment text itself.

Availability. The annotations are released under CC BY-NC 4.0 and the pipeline code under the MIT license; copyright in the underlying comment text remains with its authors. A contact point for removal requests is provided with the release.

Acknowledgments

We thank the anonymous reviewers for their careful feedback and constructive suggestions, which substantially improved this paper. We also thank Annika Blietz for her valuable contribution as the second annotator for the gold-standard dataset.

References

Dana Siobhan Atzpodien. 2022. Party competition in migration debates: The influence of the AfD on party positions in German state parliaments. *German Politics*, 31(3):381–398.

Davide Bassi, Michele Joshua Maggini, Renata Vieira, and Martín Pereira-Fariña. 2024. A pipeline for the analysis of user interactions in YouTube comments: A hybridization of LLMs and rule-based methods. In *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 146–153.

Davide Bassi, Erik Bran Marino, Renata Vieira, and Martin Pereira. 2025. Old but gold: LLM-based features and shallow learning methods for fine-grained controversy analysis in YouTube comments. In *Proceedings of the 12th Argument mining Workshop*, pages 46–57, Vienna, Austria. Association for Computational Linguistics.

Burint Bevis, Juliana Schroeder, and Michael Yeomans. 2026. Spoken disagreement is more constructive than written disagreement. *Nature Communications*, 17(1):5792.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Kai Chen, Zihao He, Rong-Ching Chang, Jonathan May, and Kristina Lerman. 2023. Anger breeds controversy: analyzing controversy and emotions on Reddit. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 44–53. Springer.

Hilde Coffé, Sam Crawley, and Josh Givens. 2026. Growing polarisation: ideology and attitudes towards climate change. *West European Politics*, 49(1):1–29.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on Twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

Paul Drecker. 2025. Dynamics of electoral polarisation in climate policy discourse: A temporal network analysis. *Politics and Governance*, 13.

Neele Falk. 2025. *What makes discussions constructive?: modeling argument and deliberative quality*. Ph.D. thesis, Universität Stuttgart, Stuttgart.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Sally Jackson. 1992. "virtual standpoints" and the pragmatics of conversational argument. In Frans H. van Eemeren, Rob Grootendorst, J. Anthony Blair, and Charles A. Willard, editors, *Argumentation Illuminated*, pages 260–269. SICSAT, Amsterdam.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. On limitations of LLM as annotator for low resource languages. *Preprint*, arXiv:2411.17637.

Julia Jakob, Timo Dobbrick, Rainer Freudenthaler, Patrik Haffner, and Hartmut Wessler. 2023. Is constructive engagement online a lost cause? toxic outrage in online user comments across democratic political systems and discussion arenas. *Communication Research*, 50(4):508–531.

Julia Jose, Meghna Manoj Nair, and Rachel Greenstadt. 2026. Large-scale analysis of persuasive content on moltbook. *Preprint*, arXiv:2603.18349.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. *Preprint*, arXiv:2004.05476.

Eugenio Levi, Michael Bayerlein, Gianluca Grimalda, and Tommaso Reggiani. 2025. Narratives of migration and political polarization: Private preferences, public preferences and social media. Working paper.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Adrian Rauchfleisch and Jonas Kaiser. 2020. The German far-right on YouTube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, 64(3):373–396.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 131–141, New York, NY, USA. Association for Computing Machinery.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Jennifer Schumann and Steve Oswald. 2024. Pragmatic perspectives on disagreement. *Journal of Language Aggression and Conflict*, 12(1):1–16.

Ozgur Can Seckin, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2026. Identifying constructive conflict in online discussions through controversial yet toxicity resilient posts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 2079–2093.

Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, 2(1):109–134.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *Preprint*, arXiv:2003.08385.

Haiyang Wang, Ye Wang, Xin Song, Bin Zhou, Xuechen Zhao, and Feng Xie. 2023. Quantifying controversy from stance, sentiment, offensiveness and sarcasm: a fine-grained controversy intensity measurement framework on a Chinese dataset. *World Wide Web*, 26(5):3607–3632.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop on the Identification of Offensive Language*, pages 1–10, Vienna, Austria. Austrian Academy of Sciences Press.

Julia Zilles and Stine Marg. 2023. Protest and polarisation in the context of energy transition and climate policy in Germany: Mindsets and collective identities. *German Politics*, 32(3):495–516.

Joseph P. Zompetti. 2015. *Divisive Discourse: The Extreme Rhetoric of Contemporary American Politics*. Cognella Academic Publishing, San Diego, CA.

A Supplemental Material: Corpus, Guidelines, and Validation

This appendix provides an overview of the German partisan corpus composition, a condensed version of the system-prompt instructions and representative gold-standard examples, and the validation matrices for our inter-annotator agreement (IAA) trials.

A.1 Corpus Demographics and Sampling Coverage

The per-channel composition of the topic-relevant corpus is presented in Table 14.

To ensure balanced representation across the discourse zones, a stratified sampling strategy was applied to generate the final 7,277-pair corpus. Table 15 details the final annotation coverage across each of the four target party-topic strata, partitioning them into human-adjudicated gold standard pairs and automated LLM-annotated pairs.

Table 16 lists the most frequent propaganda techniques across the corpus as a whole; the per-stratum breakdown for all 14 techniques is given in Table 11.

A.2 Topic-Assignment Validation

One author re-judged a random sample of 75 videos (30 climate, 30 migration, 15 other), drawn with a fixed seed, while blind to both the assigned category and the BERTopic keywords and seeing only the video title, channel, and publication date. Each video was marked as substantially about climate, about migration, about both, or about neither.

Of the 60 assigned videos, 75.0% were judged substantially on-topic (70.0% for climate, 80.0% for migration), and one of the 15 other videos was judged to contain climate or migration content. The errors are systematic rather than random: climate false positives concentrate in adjacent transport and energy debates (rail policy, fuel and gas prices), and migration false positives in general party-political content.

Because stance is annotated toward the domain, comments under an off-topic video cannot take a domain stance and are labeled Neutral. Videos assigned to a domain but judged off-topic carry 80.2% Neutral child stance against 55.0% for on-topic videos, and contributed none of the 27 opposing-stance pairs the sample produced. Topic-assignment error therefore inflates the neutral mass reported in Section 5.1 rather than contaminating the stance-relation analysis.

A.3 Annotation Guidelines and System Prompts

This section documents the formal system instructions provided to our annotators and prompt pipelines.

A.3.1 Interaction Quality Categories and Coding Rules

The annotator classifies how a child reply relates to its parent, using the five interaction-quality categories defined in Table 3 (Section 4.1).

To ensure coding stability, annotators must adhere to six operational rules. First, annotators must classify the interaction relationship between the parent and child comments rather than evaluating comments in isolation. Second, the evaluation must

Table 14: Comprehensive corpus composition by political channel and topic.

Topic	Party	Channel	Videos	Comments	Threads	Pairs
Climate	AfD	AfD TV	41	12,744	8,416	3,837
Climate	AfD	AfD-Fraktion Bundestag	345	108,034	79,164	26,279
Climate	Linke	Die Linke im Bundestag	17	1,975	1,045	777
Climate	Linke	DieLinke	41	3,162	1,767	1,179
Migration	AfD	AfD TV	168	76,659	52,931	21,062
Migration	AfD	AfD-Fraktion Bundestag	442	303,859	230,955	65,248
Migration	Linke	Die Linke im Bundestag	25	3,171	2,015	939
Migration	Linke	DieLinke	64	16,655	9,027	6,646
Climate × AfD (subtotal)			386	120,778	87,580	30,116
Climate × Linke (subtotal)			58	5,137	2,812	1,956
Migration × AfD (subtotal)			610	380,518	283,886	86,310
Migration × Linke (subtotal)			89	19,826	11,042	7,585
Total (topic-relevant)			1,143	526,259	385,320	125,967

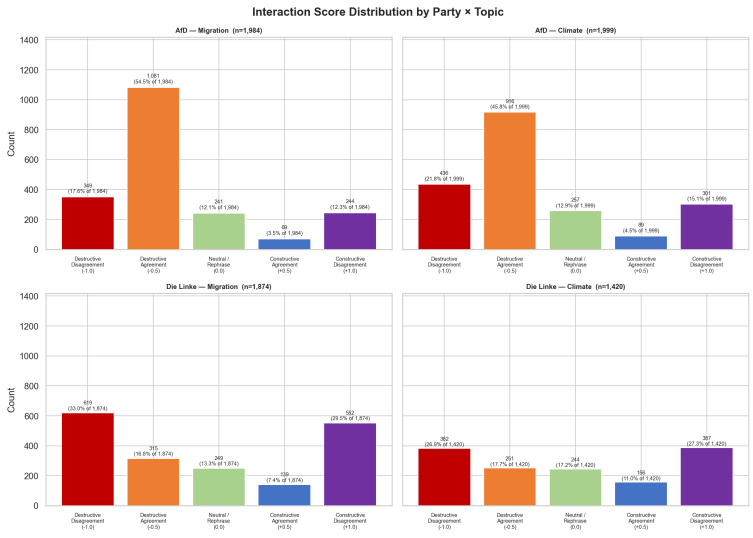


Figure 1: Distribution of interaction quality labels by party and topic.

Table 15: Annotation coverage and gold standard distribution by stratum.

Stratum	Pairs	Gold	LLM
Climate × AfD	1,999	175	1,824
Climate × Linke	1,420	175	1,245
Migration × AfD	1,984	175	1,809
Migration × Linke	1,874	175	1,699
Total Corpus	7,277	700	6,577

Note: Pairs = Gold + LLM per stratum. Gold = human-annotated (9.6% of the corpus); LLM = `claude-opus-4-6`-annotated (90.4%). Every pair carries labels on all three tasks.

Table 16: Most frequent propaganda techniques across all pairs (n = 7,277).

Rank	Propaganda Technique	Core Function	Count
1	Loaded Language	Emotive, biased phrasing	3,632
2	Name Calling / Labeling	Stereotyping and ad hominem	2,542
3	Exaggeration / Minimisation	Distorting fact severity	1,897
4	Appeal to Fear / Prejudice	Invoking collective anxiety	862
5	Causal Oversimplification	Reducing causal complexity	856
6	Doubt	Questioning credibility	837
7	Slogans	Thought-terminating clichés	707
8	Flag-Waving	Group/nationalist loyalty	524

focus strictly on communication style rather than ideological alignment. Third, any use of dehumanizing language automatically yields a score of -1.0. Fourth, sarcasm and irony must be evaluated by intent: playful sarcasm is classified by its underlying argument, whereas contemptuous or mocking sarcasm is treated as Destructive. Fifth, the guidelines prioritize substance over politeness, meaning that blunt or harsh arguments that address claims are marked as Constructive, while polite dismissals lacking reasoning are classified as Destructive. Sixth, *reductio ad absurdum* (extending arguments to extreme conclusions to expose logical flaws) counts as a valid argument and is classified as Constructive. All output is required as a single-line JSON structure matching the schema: `{"score": <float>, "label": <string>}`.

A.3.2 Stance Classifications

The stance detection pipeline is divided into issue-specific prompts for Climate Action and Migration. Table 17 consolidates the stance boundaries, numerical labels, and core criteria across both domains, serving as the system definitions for both child and parent stance tasks.

Table 17: Stance category classification definitions across domains.

Class	Label	Operational Criteria
Climate Action		
Against / Skeptic	0	Denies/downplays climate change, opposes renewable transition, mocks science, or critiques policy as destructive.
Neutral	1	Factual claims without clear positioning, neutral questions, or ambiguous/sarcastic text with unclear policy stance.
Support	2	Advocates climate action, renewables, or environmental protection; challenges skeptics.
Migration		
Against	0	Opposes immigration, portrays integration negatively, criticizes migrants, or demands assimilation.
Neutral	1	Factual, ambiguous, or irrelevant comments lacking clear stance; anger/sarcasm without policy stance.
Support	2	Defends immigrant rights, highlights positive integration, or criticizes border restrictions.

Parent context must be utilized to resolve ambiguous, short, or ironic child statements. Annotators are instructed to watch for explicit policy markers, such as references to *CO_2-Steuer* (CO_2 tax) or *Energiewende* (energy transition). Framing climate policy as an ideology or religion is treated as a typical signal of a skeptic stance (0). Crucially, a Neutral (1) classification indicates that the stance is unidentifiable, rather than implying that the comment is emotionally neutral. Stance prompts output a JSON format: `{"label": <int>}`.

A.3.3 Propaganda Techniques

For multi-label propaganda classification, the pipeline identifies the occurrence of 14 distinct rhetorical and manipulative devices within the comments. Table 18 maps each target propaganda technique to its operational definition, forcing an output schema of: `{"techniques": [<list of strings>]}` or an empty array `[]` if no techniques are detected.

A.3.4 Illustrative Gold-Standard Examples

We show one representative gold-standard example per task, illustrating how the guidelines apply. The prompts are instruction-only: no examples were included in the model input.

Interaction Quality (Destructive Disagreement). *Parent:* "Auf 26 kinder ist meine Tochter 1 von 4 Deutschen. Wenn ich sie von der Schule abhole, stehen da 8 Kopftücher und 2 normale. KOTZT MICH AN" *("Out of 26 children, my daughter is 1 of 4 Germans. When I pick her up from school there are 8 headscarves and 2 normal ones. IT DISGUSTS ME")*
Child: "Da freut sich Frau Weidel!" *("Ms. Weidel will be pleased!")*
Answer: `{"score": -1.0, "label": "Destructive Disagreement"}`

Climate Stance (Against/Skeptic). *Parent:* "Herr Dr. Spaniel... [text condensed] System runter gefahren wegen Stromausfall!" *("Mr. Dr. Spaniel... [text condensed] system shut down due to a power outage!")*
Child: "Nicht zu vergessen Reichweite ich fahre mit unserem Airsoft Team 400 km..." *("Not to forget range, I drive 400 km with our Airsoft team...")*
Answer: `{"label": 0}`

Propaganda (Loaded Language & Labeling). *Comment:* "Scheiß auf Integration." *("Screw integration.")*
Answer: `{"techniques":`

Table 18: Definitions and criteria of the 14 multi-label propaganda techniques.

Technique	Definition and Criteria
`Loaded_Language`	Words or phrases with strong emotional connotations meant to influence feelings directly.
`Name_Calling,Labeling`	Assigning labeling terms to discredit or praise without substantive argument.
`Repetition`	Deliberately repeating core claims or slogans within a single comment (excluding grammatical parallelism).
`Exaggeration,Minimisation`	Distorting factual reality via absolute quantifiers (*never, always, all*) when evidence is mixed.
`Appeal_to_fear-prejudice`	Creating anxiety about potential negative consequences to win support for a position.
`Flag-Waving`	Exploiting group identity (national, political, religious) to validate a claim.
`Causal_Oversimplification`	Reducing complex, multi-causal socio-political issues to a single, simplistic cause.
`Appeal_to_Authority`	Citing authority or expertise as sole validation without providing substantive arguments.
`Slogans/Thought-terminating_Cliches`	Using stock, simplistic catchphrases to discourage critical thought or end debate.
`Whataboutism,Straw_Men`	Deflecting criticism by alleging opponent hypocrisy or distorting their actual position.
`Black-and-White_Fallacy`	Explicitly framing a complex issue as a false binary.
`Bandwagon,Reductio_ad_hitlerum`	Leveraging popularity or negative historical associations to judge ideas.
`Doubt`	Undermining the credibility of sources by questioning their motives or expertise.
`Appeal_to_Time`	Substituting temporal assertions (*it is only a matter of time*) for logical proof.

```
["Loaded_Language",
"Name_Calling,Labeling"]}
```

A.4 Constructive-Thread Share by Year and Stratum

Table 19 gives the full temporal breakdown underlying the topic-level trend in Table 13: for each of the four party × topic strata and every sampled year, it reports the *constructive-thread share* – the percentage of threads whose mean interaction score falls in the Constructive bin ($\geq +0.25$; the top of the four thread-quality bins in Table 7). Many early-year and Die Linke cells rest on very few threads (see n in parentheses) and should be read with caution; the shares stabilize only in the better-sampled years (2023–2026), and from 2024 onward the AfD strata are consistently less constructive than the Die Linke strata.

A.5 Stance-Relation Interaction Quality by Party

Table 20 breaks the stance-relation analysis of Section 5.1 (Table 12) down by individual party × topic stratum, confirming that the climate pattern holds within both channels: opposing-stance climate exchanges are constructive on average on the AfD and Die Linke channels alike, while same-stance destructiveness is markedly stronger on the AfD side.

Table 19: Constructive-thread share – the percentage of threads whose mean interaction score lands in the Constructive bin ($\geq +0.25$) – by year and party × topic stratum (year of video publication), with thread counts n in parentheses. Strata abbreviated: C = Climate, M = Migration; A = AfD, L = Die Linke. "–" marks strata with no sampled threads that year; low-n cells are not reliable.

Year	C×A	C×L	M×A	M×L
2016	–	–	–	100.0 (1)
2017	–	100.0 (3)	–	50.0 (2)
2018	33.3 (3)	100.0 (3)	0.0 (10)	–
2019	19.4 (36)	66.7 (9)	4.9 (41)	50.0 (2)
2020	9.1 (66)	100.0 (2)	6.8 (73)	–
2021	19.3 (88)	43.8 (16)	9.8 (41)	33.3 (6)
2022	11.3 (142)	22.2 (9)	9.4 (53)	100.0 (3)
2023	7.0 (86)	7.0 (43)	8.5 (118)	20.0 (5)
2024	20.7 (29)	43.2 (44)	13.7 (95)	23.3 (133)
2025	11.4 (44)	30.9 (317)	21.0 (62)	28.3 (290)
2026	0.0 (6)	27.8 (54)	14.3 (7)	25.9 (58)

A.6 Inter-Annotator Agreement and Quality Validation

To ensure our validation gold standard is reliable, this section presents the agreement metrics computed across our independent annotation sample ($n = 700$) prior to final adjudication. Table 21 summarizes the exact agreement, Cohen's Kappa variants, and Krippendorff's Alpha coefficients across all four core tasks.

Table 22 reports two reliability views: (a) the cross-annotator interaction-quality confusion matrix between annotators A and Z, and (b) the LLM

Table 20: Interaction quality by parent–child stance relation, per party × topic stratum. Mean interaction score (range -1 to $+1$); C/D = % constructive / % destructive. "Opposing" = skeptic vs. supporter; "One side neutral" = exactly one side neutral. Opposing-stance exchanges are constructive on average (positive mean) only on climate, and on both party channels.

	Climate				Migration			
	AfD		Die Linke		AfD		Die Linke	
Stance relation	Mean	C/D	Mean	C/D	Mean	C/D	Mean	C/D
Opposing	$+0.04$	51/48	$+0.12$	55/45	-0.11	43/55	-0.06	45/53
Same stance	-0.31	15/71	-0.03	36/43	-0.34	12/77	-0.08	35/47
One side neutral	-0.28	19/67	-0.10	36/50	-0.25	20/62	-0.12	33/52

Table 21: Inter-annotator agreement ($n = 700$, pre-adjudication). κ: unweighted Cohen's κ (interaction also quadratic-weighted, 0.494); α: Krippendorff's (nominal); propaganda reports exact-set match and mean Jaccard.

Task	Exact	κ	α
Interaction Quality	56.9%	0.456	0.453
Stance (Child)	87.4%	0.640	0.639
Stance (Parent)	87.0%	0.666	0.665
Propaganda	19.3%	—	0.335[†]

[†] Mean Jaccard index, not Krippendorff's α (undefined for the multi-label task).

annotator (`claude-opus-4-6`) validated against the human gold standard, per propaganda label.

The two annotators differ mainly in strictness rather than in content-specific judgment. Annotator A is systematically harsher (mean interaction score -0.183 vs. -0.057), assigning Destructive Disagreement to 227 pairs against Z's 142, while Z assigns more Constructive Agreement (98 vs. 65). This offset is stable across topic (climate -0.104, migration -0.147; $p = 0.44$) and across party (AfD -0.086, Die Linke -0.166; $p = 0.15$), so it is consistent with a difference in threshold rather than a topic- or party-linked bias. Agreement between the two annotators is higher on each axis separately (communication style 63.4%, functional relation 67.9%) than on the joint five-class label (56.9%), and the most frequent single disagreement is A labeling a pair Destructive Disagreement where Z labels it Constructive Disagreement (45 cases): both agree on labeling the disagreement but differ on whether its tone is hostile or engaged. For propaganda, Z labels more techniques per pair (1.76 vs. 1.46), with the largest divergences on `Loaded_Language` and `Causal_Oversimplification` (more frequent for Z) against `Doubt` (more frequent for A), which accounts for the low exact-set agreement on that task.

Table 22: Reliability and annotation validation. (a) Cross-annotator interaction-quality confusion matrix between annotators A and Z ($n = 700$); classes DD (-1.0), DA (-0.5), N (0.0), CA ($+0.5$), CD ($+1.0$). (b) LLM annotator (`claude-opus-4-6`) vs. the human gold standard, per-label propaganda F_1 (global Hamming loss $= 0.1180$).

$A \backslash Z$	DD	DA	N	CA	CD	Total
DD	**120**	26	32	4	45	227
DA	5	**111**	13	14	4	147
N	7	24	**50**	36	4	121
CA	1	9	14	**35**	6	65
CD	9	6	34	9	**82**	140
Total	142	176	143	98	141	700

(a) Interaction-quality confusion matrix.

Technique	F_1
`Appeal_to_Authority`	0.3714
`Appeal_to_Time`	0.5753
`Appeal_to_fear-prejudice`	0.5792
`Bandwagon, Reductio_ad_hitlerum`	0.1702
`Black-and-White_Fallacy`	0.3333
`Causal_Oversimplification`	0.5472
`Doubt`	0.3874
`Exaggeration, Minimisation`	0.4906
`Flag-Waving`	0.4865
`Loaded_Language`	0.7565
`Name_Calling, Labeling`	0.7366
`Repetition`	0.3478
`Slogans/Thought-terminating_Cliches`	0.3500
`Whataboutism, Straw_Men`	0.3095
Macro F_1	0.4601

(b) `claude-opus-4-6` vs. gold per-label F_1.

Can Vision-Language Models Assess Environmental Reporting Compliance? A Benchmark for Requirement-Level Disclosure Alignment

Vaishali Aggarwal[1], Ishan Verma[1], Tanay Kumar Gupta[1], Tirthankar Dasgupta[2]
[1]TCS Research, New Delhi, India
[2]TCS Research, Kolkata, India
{aggarwal.vaishali, ishan.verma, gupta.tanay, dasgupta.tirthankar}@tcs.com

Abstract

Corporate sustainability reports communicate environmental disclosures through text, tables, charts, and infographics, making assessment against sustainability reporting standards challenging. Existing disclosure extraction systems focus on information extraction but do not determine whether disclosed evidence satisfies reporting requirements. We study requirement-level multimodal environmental disclosure alignment, a task that evaluates whether sustainability report evidence satisfies GRI 305 and ESRS E1-4 and E1-6 requirements. To support this task, we curate an expert-annotated benchmark comprising 731 disclosure-alignment instances from 43 sustainability reports. Using this benchmark, we evaluate three Vision Language Models (VLMs): GPT, Gemini, and LLaMA Maverick under three conditioning strategies: open-ended, framework-specific, and knowledge-augmented. Results show that although frontier VLMs possess substantial implicit knowledge of sustainability reporting standards, they frequently produce overgeneralized and optimistic alignment judgments. Requirement-level grounding improves alignment accuracy by 12% and yields more accurate disclosure gap identification. Our findings demonstrate the value of knowledge-grounded multimodal reasoning for scalable, interpretable, and auditable sustainability disclosure assessment.

1 Introduction

Emission disclosures are a central component of corporate sustainability reporting and are increasingly scrutinized by regulators, investors, and auditors (Aiuto et al., 2024). Organizations are expected to report emissions inventories, reduction initiatives, and climate targets in accordance with established sustainability reporting frameworks. Assessing whether reported information aligns with such frameworks, however, remains challenging because evidence is often distributed across narrative text, tables, charts, and infographics contained within lengthy sustainability reports (Zhang et al., 2025).

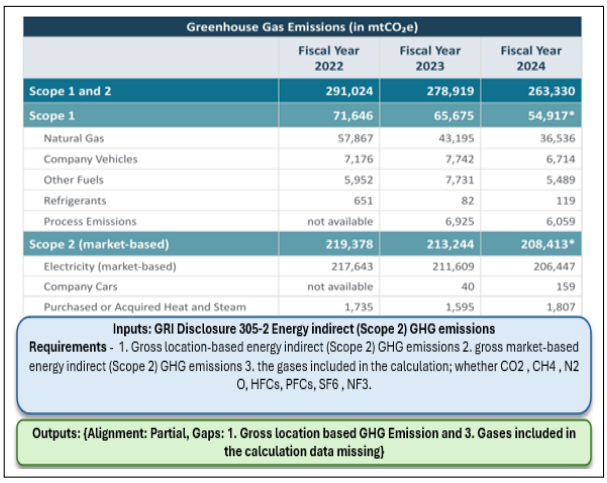


Figure 1: Example of requirement-level disclosure alignment showing supported and missing reporting requirements from multimodal report evidence.

To improve reporting consistency and comparability, several sustainability frameworks have been developed, including the Global Reporting Initiative (GRI) and the European Sustainability Reporting Standards (ESRS) (Darnall et al., 2022). Although these frameworks address similar environmental topics, they differ substantially in reporting philosophy and granularity (Usmanova et al., 2025). While GRI primarily focuses on indicator level disclosures, ESRS requires detailed methodological explanations, reporting boundaries, target-setting approaches, and assurance-ready evidence. As a result, determining whether a disclosure satisfies framework specific requirements is a non-trivial compliance task.

Recent advances in Large Language Models (LLMs) have enabled increasing research on the automated analysis of sustainability disclosures (Mishra et al., 2024; Sun et al., 2024; Us-

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 298–314
October 28, 2026 ©2026 Association for Computational Linguistics

manova and Usbeck, 2024; Goel et al., 2020). Due to the multimodal nature of sustainability reports, recent works have demonstrated the potential of VLMs for interpreting multimodal sustainability reports (Peng et al., 2024). Existing ESG (Environment, Social, and Governance) extraction systems (Anaraki et al., 2025; Wu et al., 2025; Hillebrand et al., 2023; Polignano et al., 2022; Gupta et al., 2024) identify and summarize disclosures, but do not assess whether disclosed evidence satisfies framework requirements at requirement-level granularity. However, compliance assessment requires a fundamentally different capability: determining whether disclosed evidence satisfies explicit reporting requirements. This requires grounding multimodal evidence against framework-specific criteria and identifying missing, incomplete, or weakly supported disclosures. To date, the use of Vision Language Models (VLMs) for requirement-level environmental disclosure alignment remains largely unexplored.

In this paper, we study *requirement-level multimodal environmental disclosure alignment*, a task that evaluates whether evidence contained in sustainability reports satisfies the reporting requirements of GRI 305 (Emissions) and ESRS E1-4 (Targets related to climate change mitigation and adaptation) and ESRS E1-6 (Gross Scopes 1, 2, 3 and Total GHG emissions). To support this task, we curate an expert-annotated benchmark [1] comprising 731 disclosure-alignment instances from 43 sustainability reports spanning 11 industry sectors. Using this benchmark, we evaluate the ability of frontier VLMs under progressively constrained conditioning settings: (i) open-ended setting, which relies on the model's parametric knowledge, (ii) framework-specific setting, which constrains reasoning to a particular reporting framework, and (iii) requirement-level knowledge-augmented setting, which provides the official disclosure requirements during inference. We further introduce *Coverage Breadth* and *Gap Density* metrics to analyze disclosure completeness and reporting deficiencies at the disclosure level. This work is guided by the following research questions:

- **RQ1:** To what extent can VLMs perform requirement-level environmental disclosure alignment?
- **RQ2:** How do different conditioning settings influence the accuracy and interpretability of requirement-level alignment judgments?
- **RQ3:** What role does requirement-level knowledge grounding play in reducing overgeneralization and improving disclosure-gap identification?
- **RQ4:** To what extent does requirement-level evaluation reveal reporting deficiencies that remain hidden under disclosure-level assessment?

[1] https://github.com/datasharerepo/EnvCompVLM

2 Related Work

NLP and VLMs for ESG Analysis: Recent NLP research has increasingly focused on extracting climate and sustainability related information from corporate disclosures. Prior work has explored foundation models for climate-metric extraction (Dave et al., 2024), identification of emission-reduction commitments (ClimateBERT (Schimanski et al., 2023)) and net-zero targets (NetZeroFacts (Wrzalik et al., 2024)), and climate-related QA through retrieval-augmented reasoning (Cao et al., 2025). The growing volume and complexities of sustainability disclosures have motivated LLM-based ESG analysis. Bronzini et al. (Bronzini et al., 2024) combine LLMs with RAG (Lewis et al., 2020) and in-context learning (Dong et al., 2024) to extract structured ESG insights from companies' sustainability reports for compliance assessment. ChatReport (Ni et al., 2023), ESGReveal (Zou et al., 2025), and recent LLM-based ESG extraction frameworks (Bronzini et al., 2024; Peng et al., 2024; Gutierrez-Bustamante and Espinosa-Leal, 2022) demonstrate strong capabilities for ESG information extraction and structuring. More recently, VLMs have been applied to multimodal sustainability reports (Gupta et al., 2025a; Peng et al., 2024). In particular, Wilhelmi et al. (2026) demonstrate that multimodal reasoning improves the extraction of GHG emission-reduction targets.

ESG Datasets and Benchmarks: With growing importance for ESG and sustainability analysis, several researchers have proposed different benchmark datasets: GRI-QA (Contalbo et al., 2025) for complex table QA, SustainableQA (Ali et al., 2025) for EU taxonomy QA, A3CG (Ong et al., 2025) for aspect action analysis, ClimRetrieve (Schimanski et al., 2024) and ESG-CID (Ahmed et al., 2025) for information retrieval, ESG-Bench (Sun et al., 2026) for hallucination mitigation.

Compliance & Framework Alignment: Several studies have explored sustainability framework alignment and compliance-oriented ESG analysis. sustain.AI (Hillebrand et al., 2023) and Anaraki et al. (Anaraki et al., 2025) identify report segments relevant to GRI disclosures, while ESG-KG (Chang et al., 2026) leverages standards-derived knowledge graphs to estimate compliance scores. CLAR-IESG (Santacroce et al., 2026) combines structured extraction with claim verification for ESG reporting. Gupta et al. (2025b) extract GRI disclosure data and subsequently score the extracted information based on requirements.

Collectively, prior work has advanced ESG information extraction, retrieval, QA, benchmarking, and compliance analysis. However, requirement-level disclosure satisfaction remains largely underexplored, particularly in multimodal settings using VLMs. We address this gap by introducing a benchmark for disclosure alignment and reporting gap identification, and by systematically evaluating frontier VLMs for requirement-level ESG compliance assessment.

3 Reporting Frameworks & Task Formulation

We evaluate disclosure alignment for emissions against two widely used sustainability reporting frameworks: **GRI (305)** and **ESRS (E1-4 & E1-6)**, which differ substantially in reporting philosophy and disclosure granularity.

Standard	Disclosure	Requirements
GRI 305	Direct (Scope 1) GHG Emission 305-1	1. Gross direct (Scope 1) GHG emissions 2. Gases included in the calculation; whether CO2 , CH4 , N2 O, HFCs, PFCs, SF6 , NF3 3. Biogenic CO2 emissions 4. Source of the emission factors and the global warming potential (GWP) rates used 5. consolidation approach for emissions; whether equity share, financial control, or operational control.
ESRS E1-6	P44	The undertaking shall disclose in metric tonnes of CO2eq its: (a) gross Scope 1 GHG emissions, (b) gross Scope 2 GHG emissions, (c) gross Scope 3 GHG emissions, and (d) total GHG emissions.
	P48	The disclosure on gross Scope 1 GHG emissions required shall include : (a)the gross Scope 1 GHG emissions in metric tonnes of CO2eq; and (b) the percentage of Scope 1 GHG emissions from regulated emission trading schemes.

Table 1: Example reporting requirements from GRI 305 and ESRS E1-6

GRI 305 - GRI 305 adopts an indicator-oriented reporting structure consisting of seven disclosures covering Scope 1, Scope 2, and Scope 3 greenhouse-gas (GHG) emissions, emissions intensity, emission-reduction initiatives, ozone-depleting substances, and other significant air emissions. Each disclosure is associated with a set of reporting requirements specifying the information that must be disclosed against that particular disclosure.

ESRS E1-4 & E1-6 - In contrast, ESRS E1-4 & E1-6 follow a requirement-oriented structure designed to support assurance-ready sustainability reporting under the Corporate Sustainability Reporting Directive (CSRD). The framework is organized into paragraph-level disclosures, where each paragraph contains one or more reporting requirements. These requirements cover climate targets, methodologies, reporting boundaries, Scope 1-3 emissions and emissions-trading-scheme exposure. Consequently, ESRS places greater emphasis on traceability, methodological transparency, and evidencing disclosure completeness.

Table 1 presents a few examples of how requirements are given in each standard. The complete list of GRI disclosures and ESRS paragraph-level requirements used in our experiments is provided in Appendix B. The differing structures of the two frameworks make the pair a suitable testbed for evaluating whether knowledge-grounded multimodal reasoning can identify missing, partial, and weakly supported environmental disclosures.

Task Formulation: Let f denote a reporting standard framework (GRI or ESRS), D_i^f denote the i-th disclosure under that standard, and $R_i^f = \{r_{i1}, ..., r_{im}\}$ denote the set of associated reporting requirements. Given a set $I = \{I_1, ..., I_n\}$ of extracted report pages containing evidence relevant to disclosure D_i^f, the goal is to determine whether the disclosed evidence satisfies the corresponding requirements and to identify any missing information. Formally, the VLM V predicts an alignment label A and a set of disclosure gaps G:

$$A, G = V(I, D_i^f, R_i^f)$$

where

$$A \in \{\text{Full}, \text{Partial}, \text{No}\}$$

and

$$G \subseteq R_i^f.$$

The alignment label is defined as

$$A(I, D_i^f) = \begin{cases} \text{Full} & \forall r_{ij} \in E(I, r_{ij}), \\ \text{Partial} & \exists r_{ij} \in E(I, r_{ij}), \\ \text{No} & \text{otherwise} \end{cases}$$

where $E(I, r_{ij}) \in \{0, 1\}$ denotes evidence identified in the report images corresponding to r_{ij} is sufficient or not.

$$G(I, D_i^f) = \{r_{ij} | r_{ij} \notin E(I, r_{ij})\}$$

The gap set G consists of all requirements for which sufficient supporting evidence is not found in the report images. The gap set provides an interpretable explanation of model decisions by explicitly identifying reporting requirements that lack supporting evidence.

4 Dataset Construction & VLM Conditioning Strategies

This section describes the dataset curation process to support our alignment identification task. We have also shown various VLM conditioning strategies for better alignment and gap identification for interpretable compliance check.

4.1 Benchmark Dataset Construction

Report Collection: We manually collected 56 sustainability reports from SASB repository[2]. After removing incomplete documents, and reports lacking sufficient emissions-related disclosures, the final corpus comprised 43 sustainability reports. The selected reports span all 11 SASB industry sectors, providing diverse reporting practices and levels of disclosure maturity, and environmental-impact profiles. A sector-wise distribution of companies is visualized in Figure 2.

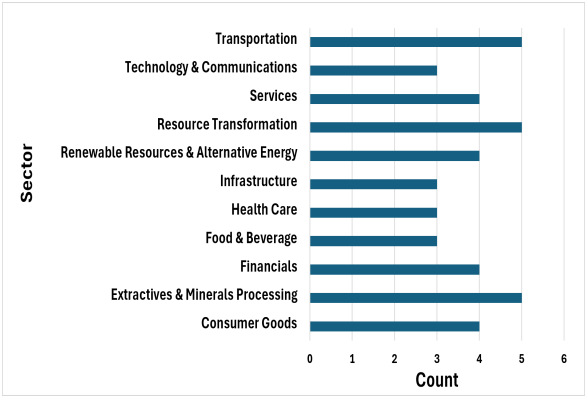


Figure 2: Company count by Sector

Evidence Page Identification: For each report, we first manually searched for GRI or ESRS content-index tables. When disclosure references were explicitly mapped to report pages, the corresponding pages were extracted directly. For reports lacking such mappings, we applied a keyword-based retrieval strategy using emission-related terms including *Scope 1*, *Scope 2*, *Scope 3*, *GHG*, *carbon emissions*, and *climate target*. Candidate pages were subsequently reviewed to ensure relevance to the selected disclosure requirements.

Requirement Construction: For each selected disclosure, the corresponding reporting requirements were manually extracted from the official framework documentation[3,4]. Each GRI disclosure was decomposed into its constituent reporting elements, while ESRS requirements were maintained at paragraph level due to their inherently granular structure. This process produced a set of disclosure-specific requirements that served as the evaluation criteria for alignment assessment.

Annotation Process: To create reference labels, three sustainability compliance experts independently reviewed the extracted report pages and assessed each requirement against the evidence present in the report. At the disclosure level, based on requirements, experts assigned one of the three alignment labels: (i) **Full**: all required information is disclosed. (ii) **Partial**: only a subset of the requirement is satisfied or evidence is incomplete. (iii) **No**: no supporting evidence for the requirement is present.

Alignment Class	Overall	GRI	ESRS
Full	269	121	148
Partial	280	124	156
No	182	56	126
Total	731	301	430
Avg. No. of Pages	3.5 pages per disclosure		

Table 2: Dataset Statistics

Disagreements were resolved through discussion and adjudication. The final annotations constitute the ground-truth alignment labels used for evaluation. The resulting benchmark contains 731 disclosure-alignment instances. Annotation reliability was measured using Fleiss' Kappa (Fleiss, 1971), yielding a score of 0.81, indicating strong inter-annotator agreement. Each benchmark instance consists of: (i) one or more report pages, (ii) a target disclosure, (iii) the associated reporting requirements and (iv) an expert alignment label.

[2]https://sasb.ifrs.org/company-use/sasb-reporters

[3]https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/

[4]https://xbrl.efrag.org/e-esrs/esrs-set1-2023.html

4.2 VLM Conditioning Strategies

We evaluate VLMs under progressively constrained evaluation strategies that move from parametric reasoning to requirement-level knowledge grounding:

- **Open-Ended Setting (Parametric Knowledge Only)**: Models receive report images and are asked to identify and align emissions-related disclosures without being informed of any specific reporting framework. No external knowledge is provided. This setting evaluates whether VLMs can rely solely on their internal knowledge to recognize and interpret environmental disclosures.

- **Framework-Specific Setting (Constrained Prompting (Yin et al., 2025))**: Models are instructed to perform alignment with respect to a specified reporting framework (GRI or ESRS), but are not provided with the corresponding disclosure requirements. This setting evaluates whether framework awareness alone improves alignment performance while still relying on the model's parametric knowledge of the framework.

- **Knowledge-Augmented Setting (Requirement Level Injection)**: In this setting, models are provided with external knowledge. Instead of relying on models' internal knowledge, models receive the report images together with the official disclosure requirements associated with the disclosures of the target framework. By grounding predictions in authoritative reporting requirements, this setting enables requirement-level assessment and reduces reliance on implicit framework knowledge. This setup reflects practical compliance-review workflows, where reporting requirements are explicitly available to analysts and auditors.

Together, these settings progressively increase the amount of framework-specific information available to the model, enabling a systematic evaluation of implicit framework knowledge, framework awareness, and requirement-level reasoning. This design allows us to assess the extent to which explicit requirement grounding improves alignment accuracy, interpretability, and disclosure-gap identification.

5 Experiments & Evaluation Metrics

5.1 Models

We evaluate three state-of-the-art Vision Language Models (VLMs): (i) **GPT 5.5** (OpenAI, 2026), (ii) **Gemini 3.5 Flash** (Google), and (iii) **LLaMA 4 Maverick** (Meta AI). These models were selected to represent both proprietary and open-weight multimodal systems with strong document-understanding capabilities. All models are evaluated under identical experimental conditions and receive the same report pages, disclosure definitions, and prompts corresponding to each evaluation setting. Before conducting the alignment experiments, we performed a preliminary capability assessment to verify two prerequisites of the proposed task: (i) comprehension of sustainability-report content, including tables, charts, and narrative text, and (ii) familiarity with reporting-framework concepts and disclosure requirements. The results indicate that all evaluated models demonstrate strong document-understanding capabilities and possess non-trivial knowledge of sustainability reporting standards. Detailed examples and analyses are provided in Appendix C.2 and C.3. Establishing these capabilities allows alignment errors to be attributed primarily to requirement-level reasoning rather than to limitations in visual perception or content extraction.

5.2 Evaluation Metrics

5.2.1 Alignment Based Metrics

The alignment task consists of three ordered labels: *No=0* < *Partial=1* < *Full=2*. Since these labels exhibit a natural ordinal structure, exact-match evaluation alone does not fully capture the severity of prediction errors. We therefore report complementary metrics that evaluate both classification performance and compliance oriented reasoning behaviour.

Accuracy: Accuracy measures the proportion of instances for which the predicted alignment label exactly matches the expert annotation. While simple and easily interpretable, Accuracy assigns the same penalty to all misclassifications regardless of their distance from the ground-truth label.

Soft Accuracy: To account for the ordinal nature of the labels, we additionally report Soft Accuracy, which assigns partial credit to predictions that differ from the ground-truth label by only one ordinal level. This metric rewards near-correct predictions while still penalizing large deviations. For

each instance, the score is defined as

$$Score_i = \begin{cases} 1.0 & \text{if } A_i^{pred} = A_i^{true}, \\ 0.5 & \text{if } |A_i^{pred} - A_i^{true}| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Soft Accuracy is then computed as

$$Soft_Acc = \frac{1}{N} \sum_{i=1}^{N} Score_i,$$

where N denotes the number of evaluated instances. Unlike Accuracy, this metric distinguishes near-miss predictions from severe classification errors.

Adjacent Error: The proportion of instances for which the predicted alignment label differs from the ground-truth label by exactly one ordinal level: $|A_i^{pred} - A_i^{true}| = 1$.

Opposite Error: The proportion of instances for which the predicted alignment label differs from the ground-truth label by two ordinal levels: $|A_i^{pred} - A_i^{true}| = 2$.

5.2.2 Analysis Based Metrics

Beyond instance-level alignment evaluation, we introduce two aggregate measures to characterize reporting completeness at the disclosure level: **Coverage Breadth (CB)** and **Gap Density (GD)**. These metrics provide complementary perspectives on disclosure completeness and enable comparisons across reporting frameworks that differ in structure and granularity. Hence, these will be computed on ground truth data.

Coverage Breadth (CB): CB measures the number of disclosures addressed across reports, regardless of whether the alignment is Full or Partial. It quantifies the proportion of reports containing evidence relevant to disclosure D_i^f.

$$CB(D_i^f) = \frac{|\{c : A_c(I, D_i^f) \in \{Full, Partial\}\}|}{N}$$

where N is the number of reports, D_i^f denotes the i^{th} disclosure within framework f, and $A_c(I, D_i^f)$ is the alignment outcome of report c for disclosure i.

Gap Density (GD): While CB captures breadth, GD represents the average proportion of unmet requirements for disclosure D_i^f across reports. It captures the extent to which reporting requirements remain unaddressed even when a disclosure is mentioned:

$$GD(D_i^f) = \frac{1}{N} \sum_{c=1}^{N} \frac{|G_c(I, D_i^f)|}{|R_i^f|}$$

where $G_c(I, D_i^f)$ denotes the set of missing requirements associated with disclosure D_i^f for report c under framework f and R_i^f is the total number of requirements associated with disclosure D_i^f.

6 Results and Discussions

In this section, we present the overall findings from the three prompting strategies and analyze disclosure-level and sector-level reporting patterns under GRI and ESRS. We further examine the relationship between alignment accuracy, Coverage Breadth (CB), and Gap Density (GD) to understand how reporting completeness influences requirement-level disclosure alignment.

6.1 Findings

We compare three conditioning strategies as mentioned in Section 4, to assess how constraint and grounding change alignment outcomes for GRI and ESRS. Table 3 reports aggregated scores across all disclosures and companies.

Open-Ended (OE): In open-ended setting, all models show limited alignment performance, particularly for ESRS. GPT achieves the strongest overall performance with a combined accuracy of 0.43 and soft accuracy of 0.60, followed closely by Gemini (0.40 and 0.49), while LLaMA performs substantially worse (0.09 and 0.17). The results suggest that frontier VLMs possess some implicit knowledge of sustainability reporting concepts, especially for the widely adopted GRI framework, where all models outperform their ESRS counterparts. This is verified by asking models "give relevant emissions disclosure and their requirements" with framework. Most models provided all disclosures and most of the requirements of the GRI. For ESRS also, they are able to give substantial information. Details are shown in Appendix C. Hence, the outputs skew toward coarse, optimistic judgments and show a bias toward GRI-style disclosures. Additionally, optimism bias remains relatively high, indicating a tendency to overestimate disclosure completeness when no explicit framework guidance is provided.

Framework Specific (FrS) Conditioning: Providing framework-specific instructions (FrS) improves performance across all models. The largest improvement is observed for LLaMA, whose combined accuracy increases from 0.09 to 0.27. GPT remains the strongest model with a combined accuracy of 0.48 and soft accuracy of 0.64. Opti-

		Gemini			Llama			GPT		
		OE	FrS	KA	OE	FrS	KA	OE	FrS	KA
GRI	**Accuracy**	0.39	0.39	0.5	0.19	0.35	0.56	0.47	0.45	0.61
	Soft Accuracy	0.51	0.53	0.68	0.29	0.47	0.72	0.62	0.63	0.75
	Adjacent Error	0.19	0.25	0.37	0.2	0.24	0.32	0.29	0.35	0.27
	Opposite Error	0.38	0.34	0.13	0.6	0.4	0.11	0.23	0.19	0.11
ESRS	**Accuracy**	0.41	0.43	0.62	0	0.19	0.69	0.39	0.5	0.67
	Soft Accuracy	0.48	0.57	0.77	0.04	0.38	0.82	0.56	0.65	0.82
	Adjacent Error	0.14	0.25	0.3	0.09	0.23	0.27	0.33	0.29	0.3
	Opposite Error	0.33	0.23	0.09	0.91	0.43	0.04	0.07	0.11	0.03
Combined	**Accuracy**	0.4	0.41	0.56	0.09	0.27	0.62	0.43	0.48	0.64
	Soft Accuracy	0.49	0.55	0.73	0.17	0.43	0.77	0.6	0.64	0.79
	Adjacent Error	0.17	0.25	0.34	0.14	0.24	0.29	0.31	0.32	0.28
	Opposite Error	0.36	0.29	0.11	0.76	0.42	0.08	0.15	0.15	0.07

Table 3: Performance of evaluated VLMs across GRI, ESRS, and combined benchmark settings under Open-Ended (OE), Framework-Specific (FrS), and Knowledge-Augmented (KA) conditioning.

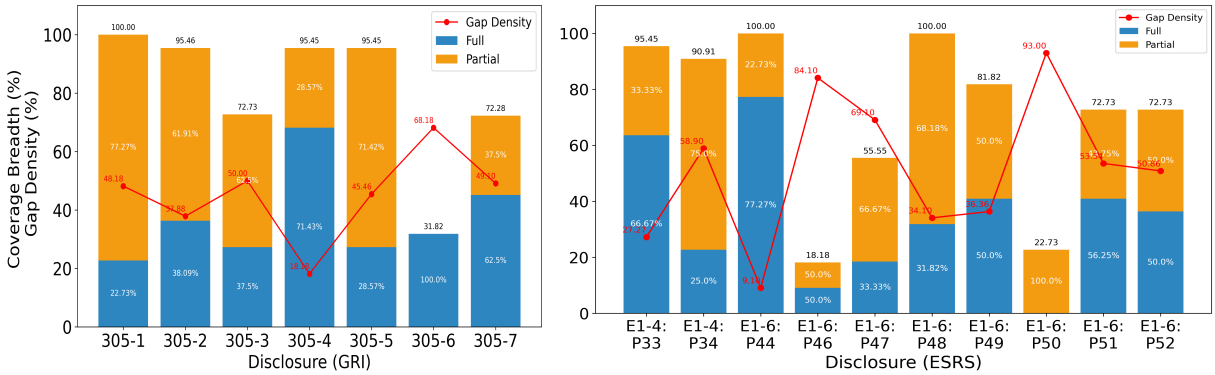


Figure 3: Disclosure wise CB(%) and GD(%) scores

mism bias also decreases across models, showing that framework guidance helps produce more disciplined alignment decisions. However, ESRS performance still lags behind GRI, indicating that framework awareness alone is insufficient for requirement-level assessment.

Knowledge Augmented (KA) Conditioning: Providing the official requirement text and asking for alignment at that granularity flips the pattern and produces the best results across all models and frameworks. GPT obtains the highest combined accuracy (0.64) and soft accuracy (0.79), while LLaMA benefits the most from requirement grounding, improving from 0.09 (OE) to 0.62 (KA). With access to the authoritative requirement text, the model's decisions became more consistent, requirement-aware, and less susceptible to over-generalisation. Performance gains are particularly pronounced for ESRS, highlighting the importance of explicit requirement-level information for complex reporting standards. Furthermore, optimism bias drops substantially for all models, suggesting that requirement grounding reduces overgeneralization and leads to more reliable compliance assessments. A key advantage of knowledge-augmented prompting is that it explains alignment decisions through requirement-level gaps. Instead of producing only an alignment label, the model explicitly identifies unmet reporting requirements, improving interpretability. Figure 1 illustrates how alignment predictions can be directly justified by the corresponding gap set. We observe that, when provided with official requirements, VLMs generate gap sets that are more consistent and closely aligned with expert annotations than those produced under open-ended or framework-specific prompting.

6.2 Disclosure-wise Analysis

Figure 3 presents Coverage Breadth (CB) and Gap Density (GD) across individual disclosures for both frameworks.

GRI- Under GRI 305, most disclosures exhibit high coverage breadth except 305-6, indicating that organizations generally address the required reporting topics. However, the corresponding gap densities reveal notable differences in reporting completeness. For example, 305-1 (Scope 1 emissions)

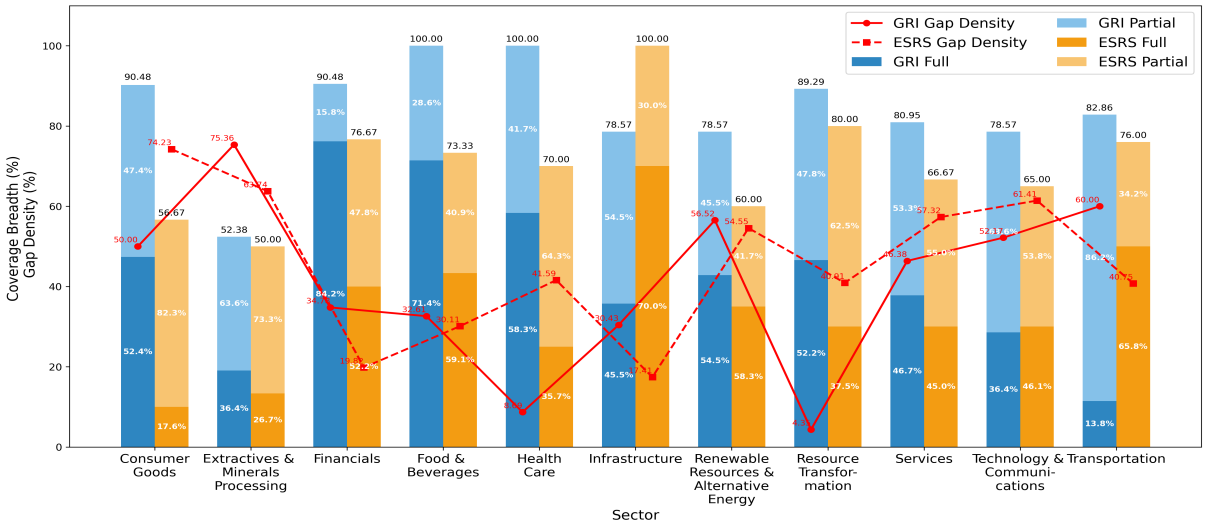


Figure 4: Sector-wise Coverage Breadth (CB%) and Gap Density (GD%) scores

and 305-5 (reduction of GHG emissions) achieve near-complete coverage (100% and 95.45%, respectively), yet maintain moderately high gap densities (48.18 and 45.46), suggesting that disclosures are often only partially complete. For 305-1, the most frequently missing requirements are *biogenic CO_2 emissions* and the *consolidation approach*. In contrast, 305-4 (GHG emissions intensity) combines high coverage breadth with the lowest gap density (18.18), making it the most consistently reported disclosure. Similar patterns are reflected in the alignment metrics (See Appendix C). 305-5 remains one of the most challenging disclosures for models despite its high coverage breadth, reflecting the difficulty of assessing reduction initiatives and supporting methodologies.

ESRS- A similar but more pronounced pattern is observed under ESRS due to its more granular and assurance-oriented requirements. Core quantitative disclosures such as P44 (Scope 1-3 and total GHG emissions) achieve complete coverage breadth (100%) and the lowest gap density (9.1), indicating both broad adoption and high reporting completeness. In contrast, disclosures related to operational control, reporting boundaries, and methodological transparency show the largest reporting deficiencies. P46 exhibits the lowest coverage breadth (18.18%) together with a very high gap density (84.1), while P47 also records high gap density despite moderate coverage. The large proportion of partially aligned disclosures for P48 and P49 further indicates that companies frequently report headline figures but omit supporting methodological details required for full compliance. Consistent with these observations, models perform best on quantitative disclosures such as P44, P49, P50, and P51 (See Appendix C). In contrast, requirements involving methodologies, assumptions, reporting-boundary changes, and comparability (P34 and P47) consistently obtain lower accuracies across all models.

6.3 Sector-wise Analysis

Sector-wise patterns as shown in Figure 4 reveal substantial differences between disclosure coverage and reporting completeness. Resource Transformation and Financials exhibit high coverage under both GRI and ESRS, but elevated ESRS gap densities indicate missing methodological details, particularly on Scope 2 emissions (E1-6:P49), boundary changes (E1-6:P47), and target disclosures. Extractives & Minerals Processing shows the weakest performance, combining low coverage with high gap densities, reflecting persistent omissions in Scope 3 emissions, pollutant data, and emissions methodologies. HealthCare performs best under GRI, achieving complete coverage with minimal reporting gaps, while Infrastructure demonstrates the strongest ESRS alignment, combining full coverage with the lowest gap density. In contrast, Consumer Goods, Services, Technology & Communications, and Transportation maintain moderate-to-high coverage but relatively high gap densities, driven by recurring omissions in pollutant emissions (GRI 305-7), significant categorywise Scope 3 reporting (ESRS E1-6:P51), target milestones (ESRS E1-6:P47), and decarbonisation-related disclosures (ESRS E1-4:P34). The sector-

wise gap patterns presented in Appendix C.4 (Figures 8a and 8b) further show that these deficiencies are concentrated around a small number of recurring disclosure requirements across sectors.

7 Conclusion

We introduced a requirement-aware formulation for multimodal environmental disclosure alignment and curated an expert-annotated benchmark covering GRI 305 and ESRS E1 requirements. Our experiments show that explicit requirement-level grounding consistently improves alignment performance across VLMs, particularly for the more granular ESRS framework. While the models exhibit some implicit knowledge of sustainability standards, parametric knowledge alone often leads to optimistic and less reliable judgments. In contrast, requirement-level gap identification provides transparent explanations for alignment decisions and supports more interpretable compliance assessment. In future work, we plan to investigate continuous requirement-level alignment scoring instead of discrete labels, develop specialized lightweight models for compliance reasoning, and extend the benchmark to additional sustainability frameworks and environmental topics. We also aim to jointly evaluate alignment quality and gap identification to support comprehensive compliance assessment.

References

Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. 2025. Enhancing retrieval for esgllm via esg-cid: A disclosure content index finetuning dataset for mapping gri and esrs. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 1–16.

Kyla Aiuto, Sarah Huckins, and Hannah Momblanco. 2024. What are greenhouse gas accounting and corporate climate disclosures? 6 questions, answered. *World Resources Institute (March 7, 2024)(www. wri. org/insights/ghg-accounting-corporate-climate-disclosures-explained)*.

Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. 2025. Sustainableqa: A comprehensive question answering dataset for corporate sustainability and eu taxonomy reporting. *arXiv preprint arXiv:2508.03000*.

Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025. 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.

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(1):41.

Zhixuan Cao, Ming Han, Jingtao Wang, and Meng Jia. 2025. Carbonchat: large language model-based corporate carbon emission analysis and climate knowledge q&a system. *arXiv preprint arXiv:2501.02031*.

Li-Yang Chang, Chih-Ming Chen, Hen-Hsen Huang, Ming-Feng Tsai, An-Zi Yen, and Chuan-Ju Wang. 2026. Esg-kg: A multi-modal knowledge graph system for automated compliance assessment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 602–608.

Michele Luca Contalbo, Sara Pederzoli, Francesco Del Buono, Venturelli Valeria, Francesco Guerra, and Matteo Paganelli. 2025. Gri-qa: a comprehensive benchmark for table question answering over environmental data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15764–15779.

Nicole Darnall, Hyunjung Ji, Kazuyuki Iwata, and Toshi H Arimura. 2022. Do esg reporting guidelines and verifications enhance firms' information disclosure? *Corporate Social Responsibility and Environmental Management*, 29(5):1214–1230.

Aditya Dave, Mengchen Zhu, Dapeng Hu, and Sachin Tiwari. 2024. Climate ai for corporate decarbonization metrics extraction. *arXiv preprint arXiv:2411.03402*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tushar Goel, Palak Jain, Ishan Verma, Lipika Dey, and Shubham Paliwal. 2020. Mining company sustainability reports to aid financial decision-making. In *Proc. of AAAI Workshop on Know. Disc. from Unstructured Data in Fin. Services*.

Google. Gemini 3.5 flash documentation. https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash. Accessed: 2026-07-16.

Tanay Gupta, Tushar Goel, and Ishan Verma. 2025a. 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.

Tanay Gupta, Tushar Goel, and Ishan Verma. 2025b. Leveraging large language models for sustainability disclosures extraction and verification. In *Australasian Conference on Information Systems (ACIS 2025)*. Association for Information Systems.

Tanay Kumar Gupta, Tushar Goel, Ishan Verma, Lipika Dey, and Sachit Bhardwaj. 2024. Knowledge graph aided llm based esg question-answering from news. In *KG4S@ ESWC*, pages 65–78.

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):9165.

Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, and 1 others. 2023. sustain. ai: a recommender system to analyze sustainability reports. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 412–416.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Meta AI. Introducing llama 4: Multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2026-07-16.

Lokesh Mishra, Cesar Berrospi, Kasper Dinkla, Diego Antognini, Francesco Fusco, Benedikt Bothur, Maksym Lysak, Nikolaos Livathinos, Ahmed Nassar, Panagiotis Vagenas, and 1 others. 2024. Esg accountability made easy: Docqa at your service. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23814–23816.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, and 1 others. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.

Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. 2025. Towards robust esg analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14854–14879.

OpenAI. 2026. Gpt-5.5 models documentation. https://developers.openai.com/api/docs/models/gpt-5.5. Accessed: 2026-07-16.

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. *arXiv preprint arXiv:2401.02992*.

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.

Marta Santacroce, Michele Luca Contalbo, Sara Pederzoli, Riccardo Benassi, Venturelli Valeria, Matteo Paganelli, and Francesco Guerra. 2026. Clariesg: An end-to-end system for esg analysis over complex tables in corporate reports. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 86–100.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524.

Siqi Sun, Ben Peng Wu, Mali Jin, Peizhen Bai, Hanpei Zhang, and Xingyi Song. 2026. Esg-bench: Benchmarking long-context esg reports for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39322–39330.

Zounachuan Sun, Ranjan Satapathy, Daixue Guo, Bo Li, Xinyuan Liu, Yangchen Zhang, Cheng-Ann Tan, Ricardo Shirota Filho, and Rick Siow Mong Goh. 2024. Information extraction: Unstructured to structured for esg reports. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 487–495. IEEE.

Aida Usmanova, Rana Abdullah, Debayan Banerjee, Markus Leippold, and Ricardo Usbeck. 2025. Reportgri: Automating gri alignment and report assessment. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6723–6727.

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.

Lars Wilhelmi, Christian Bruns, and Matthias Schumann. 2026. Enhancing the extraction of ghg emission-reduction targets from sustainability reports using vision language models. *Machine Learning and Knowledge Extraction*, 8(2):37.

Marco Wrzalik, Florian Faust, Simon Sieber, and Adrian Ulges. 2024. Netzerofacts: Two-stage emission information extraction from company reports. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 70–84.

Qilong Wu, Xiaoneng Xiang, Huang Hejia, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Ricardo Shirota Filho, and Bharadwaj Veeravalli. 2025. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1184–1203.

Xiaojie Yin, Qilong Wang, and Qinghua Hu. 2025. Constrained prompt enhancement for improving zero-shot generalization of vision-language models. *arXiv preprint arXiv:2508.17417*.

Lei Zhang, Xin Zhou, Chaoyue He, Di Wang, Yi Wu, Hong Xu, Wei Liu, and Chunyan Miao. 2025. Mmesgbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12829–12836.

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.

A Limitations

This study focuses on environmental disclosures related to GRI 305 and ESRS E1-4,6 requirements. While these frameworks cover important greenhouse-gas reporting obligations, they represent only a subset of sustainability and climate reporting standards. Consequently, the benchmark does not capture the full diversity of environmental, social, and governance disclosures encountered in practice.

The benchmark contains 731 disclosure-alignment instances collected from 43 sustainability reports spanning multiple sectors. The dataset was primarily designed as an evaluation benchmark for requirement-level disclosure alignment and compliance assessment rather than as a large-scale training resource. As a result, its size may limit the direct training of specialized models and is better suited for evaluating reasoning and compliance capabilities of existing VLMs.

Our evaluation focuses on conditioning frontier VLMs under different levels of requirement grounding rather than training task-specific models. Consequently, the findings primarily characterize the reasoning behaviour of existing VLMs when performing requirement-level compliance assessment.

Finally, although requirement-level gap identification enhances the interpretability of alignment decisions, this work primarily focuses on evaluating alignment performance and providing a qualitative analysis of disclosure gaps. Developing standardized quantitative metrics for assessing requirement-gap prediction, as well as extending the benchmark to additional sustainability reporting frameworks, remains an important avenue for future research. Furthermore, the stability and faithfulness of generated gap sets and evidence-based classification decisions warrant systematic investigation to establish the reliability and robustness of the proposed approach.

B Reporting Standards for Environmental Disclosures

GRI Framework: The GRI Standards constitute a modular and interconnected framework designed to enable organizations to publicly disclose the impacts of their activities in a structured and transparent manner. The framework comprises three series of standards, facilitating consistent and comparable sustainability reporting for stakeholders and other interested parties. For environmental disclosure, this study adopts **GRI 305: Emissions** as the relevant standard. GRI 305 encompasses a set of specific disclosures, each addressing a distinct dimension of organizational emissions.

- **GRI 305-1**: Direct (Scope 1) greenhouse gas (GHG) emissions
- **GRI 305-2**: Energy indirect (Scope 2) GHG emissions

- **GRI 305-3**: Other indirect (Scope 3) GHG emissions
- **GRI 305-4**: GHG emissions intensity
- **GRI 305-5**: Reduction of GHG emissions
- **GRI 305-6**: Emissions of ozone-depleting substances (ODS)
- **GRI 305-7**: Nitrogen oxides (NO_x), sulfur oxides (SO_x), and other significant air emissions

ESRS Framework: The European Sustainability Reporting Standards (ESRS), mandated under the Corporate Sustainability Reporting Directive (CSRD), define disclosure requirements across Environmental, Social, and Governance topics. Emission is addressed under **ESRS E1: Climate Change**. Unlike GRI's indicator-based approach, ESRS uses paragraph-level requirements for detailed reporting. For this study, we considered the following Paragraphs:

- **ESRS E1-4: P33**: Disclosure of whether/how GHG reduction targets are set to manage climate risks, impacts, and opportunities.
- **ESRS E1-4: P34**: Comprehensive GHG reduction targets: absolute and intensity values; Scope 1-3 coverage; defined base year and update cycle; 2030/2050 milestones; science-based alignment with methodologies and assumptions; and expected decarbonisation levers with their quantitative contributions.
- **ESRS E1-6: P44**: Gross Scopes 1, 2, 3 and Total GHG emissions-GHG emissions per scope
- **ESRS E1-6: P46**: Inclusion of emissions from associates/JVs/unconsolidated entities based on operational control.
- **ESRS E1-6: P47**: Disclosure of significant changes in definition of what constitutes reporting undertaking and its value chain and explanation of their effect on year-to-year comparability of reported GHG emissions.
- **ESRS E1-6: P48**: Scope 1 emissions and percentage under regulated emissions trading schemes.
- **ESRS E1-6: P49**: Scope 2 emissions reported using both location-based and market-based methods.
- **ESRS E1-6: P50**: Disaggregation of Scope 1 and 2 emissions into consolidated group vs. operational-control investees.
- **ESRS E1-6: P51**: Scope 3 emissions disclosed by each significant (material) Scope 3 category.
- **ESRS E1-6: P52**: Total emissions reported with distinction between location-based and market-based Scope 2 totals.

While GRI disclosures are often organized as indicator-style requirements, ESRS environmental requirements are more granular and assurance-oriented, which increases the risk that high-level reporting narratives appear aligned while missing required breakdowns or evidence. This contrast makes environmental disclosure alignment a suitable case for evaluating whether requirement-level knowledge grounding improves interpretability in multimodal settings and is further illustrated in Table 4, which summarises the key structural differences between GRI 305 and ESRS E1.

Aspect	GRI	ESRS
Section	GRI 305	ESRS E1-4, E1-6
Materiality	Impact Materiality (Voluntary)	Double Materiality (Mandatory under CSRD)
Audience	Broad Stakeholders	Regulators, Investors, Assurance
Targets	Not mandated in GRI 305	Targets required if set; cover Scopes 1–3, 2030/2050 milestones, science-based alignment, etc
Reduction Initiatives	Often Qualitative	Quantified levers required
Biogenic CO2	Varies by reporter	Explicit separation/treatment expected
Assurance Orientation	Low Assurance	Assurance-ready design

Table 4: Comparison of GRI and ESRS Frameworks for GHG Emissions

C Additional Results

C.1 Paragraph-Level Performance under ESRS

Table 5 reports knowledge-augmented alignment performance across individual ESRS requirements. Overall, all three models perform well on requirements centered around explicit quantitative disclosures, particularly P44 (Scope 1–3 and total emissions), P49 (location- and market-based Scope 2 emissions), P50 (disaggregation of Scope 1 and

	Gemini				GPT				Llama			
	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp
E1-4: P33	0.64	0.76	0.25	0.11	0.54	0.71	0.36	0.1	0.61	0.79	0.36	0.03
E1-4: P34	0.46	0.71	0.5	0.04	0.5	0.73	0.46	0.04	0.57	0.77	0.39	0.04
E1-6: P44	0.61	0.80	0.39	0	0.75	0.88	0.25	0	0.93	0.96	0.07	0
E1-6: P46	0.75	0.84	0.18	0.17	0.71	0.86	0.29	0	0.79	0.89	0.21	0
E1-6: P47	0.36	0.57	0.43	0.21	0.36	0.64	0.57	0.07	0.57	0.71	0.29	0.14
E1-6: P48	0.54	0.70	0.32	0.14	0.71	0.82	0.21	0.07	0.71	0.86	0.29	0
E1-6: P49	0.75	0.86	0.25	0	0.82	0.91	0.18	0	0.82	0.89	0.14	0.04
E1-6: P50	0.79	0.89	0.21	0	0.89	0.95	0.11	0	0.64	0.82	0.36	0
E1-6: P51	0.75	0.84	0.18	0.07	0.79	0.88	0.18	0.03	0.79	0.86	0.14	0.07
E1-6: P52	0.57	0.73	0.32	0.11	0.61	0.80	0.39	0	0.54	0.75	0.43	0.04

Table 5: Paragraph-wise evaluation of VLMs on ESRS requirements under the Knowledge-Augmented (KA) setting. Results are reported using Accuracy (Acc), Soft_Acc (SA), Adjacent Error (Adj), and Opposite Error (Opp).

2 emissions), and P51 (Scope 3 category-wise reporting). These requirements achieve the highest accuracies and soft accuracies across models, with LLaMA reaching 0.93 accuracy on P44 and GPT achieving 0.89 accuracy on P50. The consistently low optimism bias for these requirements suggests that VLMs can reliably align disclosures when evidence is presented in structured tables and numerical summaries.

In contrast, all models struggle more with P34 and P47, which require reasoning over methodologies, assumptions, target-setting approaches, reporting-boundary changes, and year-to-year comparability. These requirements exhibit the lowest accuracies across models, with P47 obtaining only 0.36 accuracy for Gemini and GPT. They also produce higher optimism bias and adjacency scores, indicating that models frequently identify partially relevant evidence but have difficulty determining whether all required methodological details are explicitly disclosed.

Comparing models, GPT and LLaMA generally outperform Gemini, although their strengths differ across requirement types. GPT demonstrates the most stable performance across requirements and maintains consistently low optimism bias, while LLaMA achieves the highest scores on several quantitative disclosure requirements, including P44 and P49. However, all models show a similar trend: performance is substantially higher for requirements requiring factual extraction than for those requiring verification of reporting methodology, assumptions, and disclosure completeness. This suggests that requirement-level grounding is effective for identifying explicit evidence but remains challenging for more interpretive and assurance-oriented ESRS requirements.

C.2 Disclosure-Level Performance under GRI

Table 6 presents requirement-level results for the GRI 305 disclosures under knowledge-augmented prompting. Overall, the models achieve their strongest performance on the core emissions inventory disclosures, particularly 305-1 (Scope 1 emissions) and 305-4 (emissions intensity), where GPT obtains the highest accuracies of 0.75 and 0.75, respectively. All three models also demonstrate low optimism bias for these disclosures, suggesting that explicit and quantitative emissions data can be reliably aligned with GRI requirements.

Performance declines for disclosures that require more detailed contextual information about gases is required. In particular, 305-5 (reduction of GHG emissions) is consistently more challenging across all models.

Among the evaluated models, GPT achieves the most consistent overall performance, leading on five of the seven GRI disclosures. LLaMA performs competitively on 305-5 and maintains the highest soft accuracy for this disclosure (0.82), while Gemini achieves the strongest result on 305-6 (ozone-depleting substances) with an accuracy of 0.72. Across all models, disclosures requiring direct extraction of reported emission values are substantially easier than disclosures involving reduction initiatives, Scope 3 reporting, and pollutant emissions, highlighting that methodological completeness remains the primary challenge in requirement-level GRI alignment.

Figure 5 demonstrates how outputs under the knowledge augmented prompting strategy were classified into Full, Partial, and No Alignment. **Full**

	Gemini				GPT				Llama			
	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp
305-1	0.6	0.8	0.4	0	0.75	0.875	0.25	0	0.72	0.86	0.28	0
305-2	0.4	0.7	0.6	0	0.65	0.825	0.35	0	0.52	0.76	0.48	0
305-3	0.44	0.64	0.4	0.16	0.6	0.75	0.3	0.1	0.48	0.66	0.36	0.12
305-4	0.64	0.8	0.32	0.04	0.75	0.875	0.25	0	0.6	0.8	0.4	0
305-5	0.32	0.6	0.56	0.12	0.55	0.725	0.35	0.1	0.64	0.82	0.36	0
305-6	0.72	0.72	0	0.28	0.65	0.675	0.05	0.3	0.68	0.68	0	0.32
305-7	0.4	0.56	0.32	0.28	0.4	0.575	0.35	0.25	0.28	0.46	0.36	0.36

Table 6: Disclosure-wise evaluation of VLMs on GRI requirements under the Knowledge-Augmented (KA) setting. Results are reported using Accuracy (Acc), Soft_Acc (SA), Adjacent Error (Adj), and Opposite Error (Opp).

Requirement(s)	Output	Alignment
Gases included in the calculation (Scope 1); whether CO2 , CH4 , N2 O, HFCs, PFCs, SF6 , NF3 , or all	The image states that "In scope 1, we only consider methane and N2O emissions of biogenic emissions, following SBTi guidance, amounting to 2 t of CO2e in 2024.	Full
	The second image explains that GHG emissions are calculated in accordance with the GHG Protocol, which includes various gases.	Partial
	The images do not directly provide the gross Scope 3 emissions. However, the tables on page 11 and 12 list various GHG emissions disclosures, including Scope 1 and Scope 2 emissions, but do not explicitly mention Scope 3 emissions.	No

Figure 5: Example: An alignment based on requirement-level evidence for disclosure 305-1

alignment indicates that all parts of the requirement are explicitly satisfied. **Partial alignment** applies when only some components of the requirement are met or when references are indirect. **No Alignment** is assigned when the response lacks any evidence relevant to the requirement. This categorization ensures fine-grained evaluation of compliance at the requirement level.

Models Internal Knowledge about GRI and ESRS: For GRI, all models generated most of requirements. While Gemini generated 96%, GPT generated 93% and Llama generated 85% of the requirements. Showing higher knowledge corresponding towards GRI. In contrast, for ESRS all models generated weaker paragraphs. Exact numbers can not be calculated for ESRS because of their paragraphs. While qualitative results suggest Gemini appears to possess the most relevant knowledge, followed by GPT. While Llama has the lowest exact knowledge.

Task Complexity: Aligning report content with Sustainability disclosure frameworks is inherently complex. Sustainability reports are typically unstructured, lengthy, and multimodal, containing a mix of narrative text, quantitative tables, charts, and infographics. Unlike conventional structured benchmarks, the content spans heterogeneous industries and employs diverse terminologies, making it challenging for models to consistently identify relevant disclosures. Further, different frameworks impose requirements at varying levels of granularity and domain specificity. Although many disclosures partially overlap across frameworks, they differ in scope, emphasis, and interpretive expectations, making direct model-driven disclosure alignment a non-trivial interpretability challenge. We analyze alignment behaviour across progressively constrained prompting regimes, examining how varying degrees of guidance and knowledge grounding influence the model's ability to produce requirement-aware judgments and surface weakly supported or incomplete disclosures. Our objective is to advance AI-assisted, auditable assessment of environmental reporting by clarifying how prompting design influences interpretability and by providing aggregate indicators of reporting coverage introduced in Section 4.

C.3 Baseline Content Understanding

This section provides an example used to verify that the evaluated VLMs can accurately interpret sustainability-report visuals before performing disclosure alignment. The sample includes the original report image (Figure 6) and the corresponding raw output generated by the model (Figure 7). The multi-modality of disclosures can be seen in the report page images. No framework instructions or disclosure requirements were provided; the model was only expected to extract and summarize the information present in the image. This example confirms that the subsequent challenges in disclosure alignment arise from framework-level reasoning rather than limitations in visual or textual comprehension.

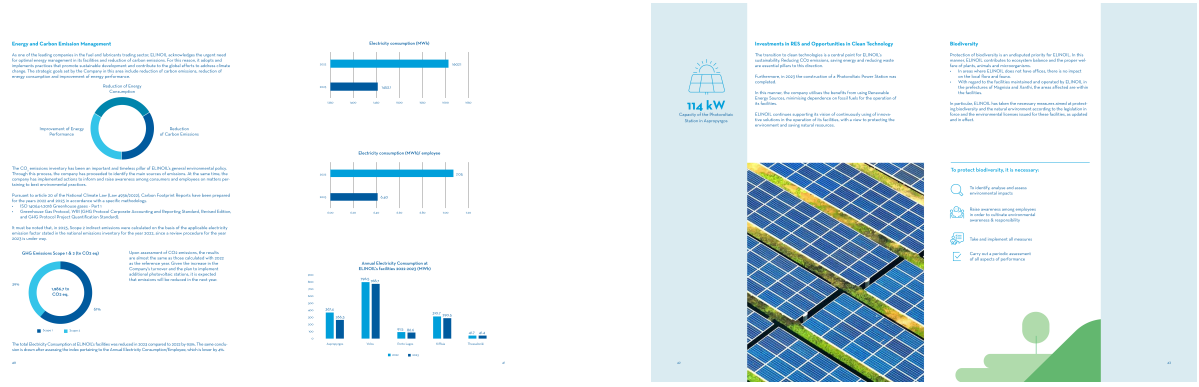


Figure 6: Images from a sustainability report

Prompt:
You are given multiple images from a company's sustainability report. Describe the content provided in the images in detail and image-wise. Be specific with the sections and the related content. Also analyze the infographics, tables, chart, etc. in the images, if present.
Raw Output:
The provided images appear to be from a company's sustainability report, specifically focusing on energy and carbon emission management, investments in renewable energy sources (RES) and clean technology, and biodiversity. Here's a detailed description of the content in each image: **Image 1: Energy and Carbon Emission Management** * The section begins with a description of XYZ's commitment to optimal energy management and reduction of carbon emissions. * A circular diagram illustrates the company's strategic goals: * Reduction of Energy Consumption * Improvement of Energy Performance * Reduction of Carbon Emissions * The CO2 emissions inventory is a crucial part of XYZ's environmental policy, aiming to identify main emission sources and promote best environmental practices. * A pie chart shows the distribution of GHG Emissions Scope 1 and 2 (tn CO2 eq): * Scope 1: 39% * Scope 2: 61% * The total electricity consumption at XYZ's facilities decreased by 9.5% in 2023 compared to 2022. * Three bar charts display: * Electricity consumption (MWh) for 2022 and 2023 * 2022: 1607.1 MWh * 2023: 1453.1 MWh * Electricity consumption (MWh)/employee for 2022 and 2023 * 2022: 7.05 MWh/employee * 2023: 6.40 MWh/employee * Annual Electricity Consumption at XYZ's facilities 2022-2023 (MWh) * Aspropyrgos: 367.4 (2022), 266.3 (2023) * Volos: 796.5 (2022), 768.7 (2023) * Porto Lagos: 91.5 (2022), 86.6 (2023) * Kiffisia: 310.7 (2022), 290.5 (2023) * Thessaloniki: 41.7 (2022), 41.4 (2023) **Image 2: Investments in RES and Opportunities in Clean Technology and Biodiversity** * The section highlights XYZ's investments in RES and clean technology, including: * A photovoltaic station in Aspropyrgos with a capacity of 114 kW * The construction of a Photovoltaic Power Station in 2023 * The company's vision for using innovative solutions to protect the environment and save natural resources is emphasized. * A section on biodiversity discusses XYZ's efforts to: * Protect ecosystem balance and welfare of plants, animals, and microorganisms * Identify, analyze, and assess environmental impacts * Raise awareness among employees to cultivate environmental awareness and responsibility * Take and implement measures to protect biodiversity * Carry out periodic assessments of performance The infographics, tables, and charts in the images effectively illustrate XYZ's progress and goals in energy and carbon emission management, investments in RES and clean technology, and biodiversity conservation. The data presented demonstrates the company's commitment to sustainability and environmental responsibility.

Figure 7: Sample prompt used with model output

C.4 GRI vs ESRS Findings and Compliance Critique

For the same set of reports, coverage breadth is steadier and generally higher under GRI because core requirements are widely disclosed, whereas ESRS breadth is lower and more variable because coverage depends on paragraph-level evidence that many sectors only partially provide.

Gap	Sectors										
	Consumer Goods	Extractives & Mineral Processing	Financials	Food & Beverage	Health Care	Infrastructure	Renewable Resources & Alternative Energy	Resource Transform-ation	Services	Technology & Communic-ations	Transport-ation
Scope 3 emissions incomplete/unclear	☑	☑	☐	☐	☑	☐	☐	☐	☐	☑	☐
Biogenic CO_2 missing	☑	☑	☑	☑	☑	☐	☐	☑	☑	☑	☑
Reduction Initiatives lack quantification	☐	☐	☐	☑	☐	☐	☐	☑	☑	☐	☑
Consolidation approach unclear	☑	☑	☑	☑	☐	☑	☑	☐	☑	☑	☐
Pollutant Data Missing	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑

(a) GRI 305

Gap	Sectors										
	Consumer Goods	Extractives & Mineral Processing	Financials	Food & Beverage	Health Care	Infrastructure	Renewable Resources & Alternative Energy	Resource Transform-ation	Services	Technology & Communic-ations	Transport-ation
Scope 3 not disclosed by each significant category	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Scope 1/2 not disaggregated	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Associates/JVs/unconsolidated entities not included by operational control	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Scope 1: % of emissions under ETS not disclosed	☑	☑	☐	☑	☐	☑	☐	☑	☑	☐	☐
Scope 2: location-based and market-based emissions both not provided	☑	☑	☐	☑	☑	☐	☐	☑	☑	☑	☐
Total GHG not disclosed as dual totals	☑	☑	☐	☑	☑	☑	☐	☑	☑	☑	☐
Boundary changes not explained	☑	☑	☑	☑	☑	☑	☐	☑	☑	☑	☐
Reduction targets are set not disclosed / unclear	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Targets lack absolute/intensity basis or	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Targets missing required milestones (≥2030; if available 2050)	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Target methodology & critical assumptions	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Decarbonisation levers not described with quantitative contributions	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑

(b) ESRS E1-4, E1-6

Figure 8: Sector-wise gaps across reporting framework disclosures

The common gaps (Figure 8) that appear repeatedly across sectors are *Significant Scope 3 details*, *boundary/operational-control clarity*, *Location-based/Market-based handling with dual totals*, *Emissions Trading System share*, and *auditable target mechanics*. These are precisely those that ESRS requires to be explicit, and their recurrence explains the higher ESRS gap density seen in Figure 4. Practically, the sector path to improvement is to raise ESRS coverage while shrinking those recurring gaps, so that breadth converts into assurance-ready depth rather than remaining as partial alignment with persistent methodological deficits.

C.5 Prompt Templates

For reproducibility, we provide the prompts used (Figure 9, 10, 11) under the three conditional settings: Open-Ended (OE), Framework-Specific (FrS), and Knowledge-Augmented (KA). Deterministic decoding was employed to ensure reproducibility by eliminating stochasticity during text generation and consistently selecting the most probable token at each decoding step.

Open Ended Prompt
You are a sustainability reporting analyst. Given one or more images from a company's sustainability report, identify whether the content aligns with GRI and/or ESRS emission-related disclosures based on explicit or implicit evidence in the report, disclosures, and reporting requirements reflected in the content. For GRI, provide the disclosure-wise analysis while for ESRS, provide the paragraph-wise alignment. For each identified disclosure or requirement: • Provide the disclosure/requirement code and title. • Determine whether it is Fully Aligned, Partially Aligned, or Not Aligned. For partially aligned or non-aligned disclosures, identify the missing information required for full compliance. Consider all available report content, including text, tables, charts, and visual elements, and use official disclosure or requirement identifiers whenever possible.

Figure 9: Prompt used for Open-ended setting

Framework Specific Prompt
You are a sustainability reporting analyst. Given one or more images from a company's sustainability report, assess alignment with the specified sustainability reporting framework (GRI/ESRS). Instructions: 1. Identify the emission-related disclosures or requirements addressed in the images. 2. For each disclosure/requirement: • Provide the official disclosure/requirement code and title. • Determine whether it is Fully Aligned, Partially Aligned, or Not Aligned. • Cite the supporting image(s) and briefly justify the assessment. 3. Identify any missing information required for disclosures or requirements that are partially or not aligned. 4. Provide an overall alignment assessment for the framework. Output Format: Disclosure/Requirement: <code> - <title> • Alignment: • Supporting Image(s): • Rationale: Overall Alignment: Missing Elements:

Figure 10: Prompt used for Framework-specific setting

Knowledge Augmented Prompt
You are a sustainability reporting analyst. You are provided with one or more images from a company's sustainability report and a disclosure along with its reporting requirements for GRI (or ESRS) framework. Disclosure: {disclosure} Disclosure Name: {disclosure_name} Requirements: {requirements} Assess how well the disclosure requirements are satisfied based on all provided images. Instructions: 1. Evaluate each requirement independently. 2. For each requirement, determine whether it is Fully Addressed, Partially Addressed or Not Addressed 3. Based on all requirement-level assessments, determine the overall disclosure alignment as Fully Aligned, Partially Aligned or Not Aligned Output Format: Req 1 - <requirement description> - <alignment status> Req 2 - <requirement description> - <alignment status>... Overall Alignment - <overall alignment status> Return only the required output in the specified format. Do not provide explanations, reasoning, or additional information.

Figure 11: Prompt used for Knowledge-Augmented setting

Proportionate Responses: The Case for Small Language Models as Sustainable Expert Climate Policy Annotators

Jetske Bonenkamp, Robbert Biesbroek, Ioannis N. Athanasiadis
Wageningen University and Research, Wageningen, the Netherlands
(jetske.bonenkamp, robbert.biesbroek, ioannis.athanasiadis)@wur.nl

Abstract

Tracking progress on climate change adaptation requires systematic and robust annotation of adaptation policies, but manual annotation has become infeasible due to the rapidly growing volume of unstructured documents. Pre-trained language models (LMs) have emerged as promising tools for automating and scaling annotation. However, relatively little is known about their accuracy and quality on complex, free-text annotation tasks. Meanwhile, large LMs (LLMs) are increasingly deployed for large scale text analyses despite their considerable environmental footprint and limited exploration of more sustainable alternatives. By benchmarking model accuracy against human inter-annotator agreement, reviewing perceived annotation quality, and estimating energy and water consumption, we show that fine-tuning small LMs (SLMs) provides a substantially more sustainable alternative to LLMs with only marginal reductions in accuracy and improved perceived annotation quality. Furthermore, we demonstrate the importance of evaluating complex annotation tasks beyond accuracy against human gold standard annotation as ambiguity and human error can obscure meaningful differences in model performance. By reducing the costs of high-quality policy annotation, our approach enables the scalable knowledge extraction needed for tracking progress on adaptation, living evidence syntheses, and evidence-informed policy evaluation.

1 Introduction

Global reporting requirements resulted in a highly voluminous, continously expanding landscape of climate policy documents. Since the Paris Agreement and its Global Goal on Adaptation (GGA), such documents increasingly cover adaptation *goals* (i.e., statements of government aims and ambitions that define the overall direction of policy) and *instruments* (i.e., tools or resources through which governments seek to achieve their goals). These goals and instruments, referred to as "policy elements", form crucial evidence for tracking adaptation progress - an issue that has become increasingly urgent in climate debates, particularly following the recently adopted set of 59 indicators to track the targets under the United Arab Emirates Framework for Global Climate Resilience (Rubial et al., 2023; UNFCCC, 2025). These indicators include the adoption and implementation of national adaptation plans, policies, and strategies. The unstructured nature of the evidence (Bonenkamp et al., 2025), combined with the complexity and ambiguity of the concept of adaptation (Orlove, 2022), however, complicates adaptation tracking and has resulted in an incomplete understanding of the current state of adaptation globally. To address this gap, establishing a structured, global-coverage database of adaptation policy at the element level plays a critical role.

Transforming unstructured, multi-sectoral documents into granular, usable evidence traditionally relies on expert text annotation. However, with databases such as Climate Policy Radar already containing over 16,000 climate policy documents published between 2010 and 2025, while the number of documents increases rapidly, manual annotation has become infeasible. Consequently, research has increasingly turned to automated approaches for data processing and analysis (Bonenkamp et al., 2025; Sietsma et al., 2024). Recent advances in the field of Natural Language Processing (NLP), particularly the rise of pre-trained language models, have enabled moving beyond traditional approaches (e.g., text-as-data through topic modeling or text classification) and allow for automating complex text tasks, including detailed annotation processes. In recent years, the use of general-purpose Large Language Model (LLM)s became increasingly common in diverse research settings, given their highly advanced capabilities in many tasks without the need for training data (Wan et al.,

Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026), pages 315–340
October 28, 2026 ©2026 Association for Computational Linguistics

2024). Meanwhile, the high computational demand and, consequently, environmental cost of these large models is often not aligned with their performance gains over solutions with smaller, domain-specific LMs - particularly in big data applications (Wang et al., 2025).

The trade-off between the environmental sustainability of annotation models and the accuracy and quality of automated annotations is difficult to determine in a complex and ambiguous task like annotating adaptation policies. For evaluating the accuracy of NLP models, it is common to rely on hand-labelled "gold standard" data. Expert annotation of free text is inherently complex, granular, and ambiguous, resulting in only moderate inter-annotator agreement (IAA) (Kasner et al., 2025). Consequently, evaluating automated annotation is less straightforward than in, e.g., binary text classification tasks, and a human-labelled gold standard alone may not fully reflect annotation quality.

In this work, we take a critical view on the use of LLMs for large-scale text analysis and assess the potential of Small Language Model (SLM)s as expert climate policy annotators. We consider generative LMs with less than one billion parameters, as well as encoder-only LMs (e.g., BERT models) as SLMs. Although prior work has benchmarked accuracy and computational cost of SLMs for broad NLP tasks with general-purpose (Pham et al., 2025) or fine-tuned (Garg et al., 2025) SLMs, the evaluated NLP tasks are not comparable to the policy annotation task in terms of complexity and domain specificity. As a result, a research gap around the potential of *specialist* SLMs as free-text annotators, particularly in the climate and/or policy domain, has persisted.

To fill this gap, we introduce a framework of fine-tuned, specialist SLMs to extract and categorize adaptation policy elements that have remained hidden thus far. The annotation process is divided into six tasks, each performed by a specific SLM. We benchmark the fine-tuned SLMs against IAA metrics and annotations generated by LLMs across three dimensions: annotation *accuracy*, annotation *quality*, and environmental *sustainability*. In doing so, we assess whether SLMs provide a viable and sustainable approach for upscaling adaptation policy annotation and building a global-coverage database of structured adaptation policy evidence. To assess annotation *accuracy*, we compare model-human and human-human F1 scores while accounting for partial overlap and adaptation intentionality. The assessment of *quality* uses experts' perceptions and preferences of model- and human-created annotations, determined through a blind review process. Finally, environmental *sustainability* is assessed by estimating the total energy use, carbon footprint, and water consumption that each model would require for processing global-coverage data.

2 Manual annotation

To allow for fine-tuning SLMs, estimate IAA, and measure model accuracy, we created a hand-labelled dataset. First, based on several existing, related typologies, we defined a detailed code book for annotating adaptation policy elements - the full coding scheme is provided in Appendix A. A policy element is based on a text span - a verbatim sequence of words extracted from the original text chunk - and can be assigned type *Goal* or *Instrument*. Within these element types, there are different subclasses (five for *Goal*, six for *Instrument*; see Appendix Table 1). Furthermore, each element can be assigned a hazard and/or sector class (where applicable) out of five classes each, indicating the climatic hazard the element addresses and the sector it targets (see Appendix Table 2). In addition, an adaptation intentionality score is assigned: this 10-point score (0-1) indicates the extent to which the policy element intentionally contributes to increasing resilience (IPCC, 2022) as opposed to indirectly.

Following the code book and building upon a database of pre-classified adaptation-relevant text chunks that average 10 paragraphs (see Bonenkamp et al. (2025) for the details), we extracted a stratified sample of 2,644 chunks used for manual annotation. A team of four trained, graduate-level annotators spent over 600 hours annotating the dataset, including training and iterative refinement of the code book[1]. This resulted in identification and categorization of over 12,000 adaptation policy elements. From this hand-labelled dataset, we extract stratified samples: 2,453 (training) and 191 (testing) input chunks and 11,890 (training) and 847 (testing) output policy elements. Among the extracted policy elements, 69% is labelled as *Goal*; 31% as type *Instrument*.

[1]The annotation process was conducted by research assistants who were compensated in accordance with standard institutional pay scales and local labor law requirements.

3 Model selection

The annotation framework consists of three specialist models that are all fine-tuned[2] on the hand-labelled data introduced in section 2. The specialist models (Table 1, column C) complement each other to extract and categorize adaptation policy evidence and are responsible for six different tasks in the process: extracting the policy elements, assigning (sub)types, hazards, and sectors, and predicting adaptation intentionality (Table 1, column A). For each task, the specialist models are benchmarked against human IAA and one or more alternative models (Table 1, column D). For all tasks, we use OpenAI's *gpt-5-mini* model as a benchmark LLM and prompt it to complete the task while providing the full code book descriptions, but no examples[3]. For the second task (classifying policy elements as *Goal* or *Instrument*), we additionally benchmark a fine-tuned *ClimateBERT* model to compare the accuracy of the SLM in directly predicting the element type during element extraction.

Considering that the first task (extracting the element text spans) involves language generation, we use a generative SLM for this task. To capture the differences between goals and instruments directly, the model is fine-tuned to return objects containing both the element span and the applicable element type, combining the first two tasks at once (Table 1). For the first two tasks, we selected the 0.6 billion parameter version of *Qwen 3* (Yang et al., 2025) as our base model to demonstrate the potential of models with relatively few parameters. This advanced, recent SLM allows full fine-tuning, while required computational resources remain limited. We fine-tune the model to take the full chunks as input and return JSON objects (see Appendix D for finetuning details) containing the policy element text spans and their element type. We refer to this first SLM as the **annotation model**. For evaluating its accuracy in predicting the right element type, we benchmark the model against the predictions of both *gpt-5-mini* and a fine-tuned *ClimateBERT* model (Webersinke et al., 2022)

ClimateBERT is a language model that is pretrained on textual data from the climate domain. In prior, related work, this model has shown performance gains over similar models that are not pretrained on climate-related data (Bonenkamp et al., 2025). As BERT-like models are encoder-only, they are appropriate for tasks that do not involve language generation. Therefore, we use *ClimateBERT* as the base model for our classification and regression tasks and fine-tune two separate models. The first *ClimateBERT* model, the **policy NLI classifier**, covers tasks 3 up to 5. As its name suggests, this model is fine-tuned on data inspired by Natural Language Inference (NLI): it predicts whether the policy element (premise) belongs to the class description (hypothesis). A set of 48,000 premise-hypothesis pairs is used for fine-tuning, comprising sequence-class description pairs with the ground truth class and an *entailment* label (1), as well as sequences paired with an incorrect class, on which the model learns when to predict the *neutral* label (0).

The NLI dataset was composed by adding each policy element in the original training dataset once with its correct subtype label (entailment), twice with an incorrect subtype label (neutral), with its correct hazard and/or sector where applicable (entailment), and a random sample of 5,000 elements that were not assigned a hazard or sector (neutral). The class descriptions used for the *neutral* pairs are randomly selected based on the dataset's class distribution. A few examples of the premise-hypothesis pairs are provided in Appendix Table 5 (Appendix C). This single-model NLI approach was used to optimize performance on classification of underrepresented classes, given the highly imbalanced training dataset (see Appendix B). Rather than learning class-specific decision boundaries alone, the model learns to recognize patterns between a policy element and class descriptions, enabling better generalization to classes with relatively few training examples.

For the final task (predicting adaptation intentionality), we fine-tuned another *ClimateBERT* model on a simple regression task, using standardized intentionality scores assigned by the human annotators. This model, the **intentionality predictor**, assigns an adaptation-intentionality score based on the policy element as input. This task can be seen as both a regression, as well as an ordinal classification task: the score can be used to group elements into intentionality classes for policy analyses. For the benchmark approach with *gpt-5-mini*, we therefore map the problem onto three ordered categories (*indirect*, *implicit*, and *intentional* adaptation) and align the LLM-predicted categories to

[2]All models in the framework are fine-tuned end-to-end. Fine-tuning details are provided in Appendix D.

[3]A few-shot approach was tested during the initial research phase, showing a substantially increased inference time while accuracy did not improve.

	A	B	C	D
#	**Task name**	**Description**	**Specialist model**	**Benchmark model(s)**
1	Element extraction	Extracting policy element text span	**Annotation model (SLM 1)** Fine-tuned *Qwen3-0.6B* Input: adaptation text chunk (*n=2,453*) Output: list of JSON objects (*n=11,890*)	gpt-5-mini (JSON mode)
2	Type classification	Assigning element type (*Goal* or *Instrument*)		ClimateBERT, gpt-5-mini
3	Subtype classification	Predicting subtype (e.g., process-oriented goal)	**Policy NLI classifier (SLM 2)** Fine-tuned *ClimateBERT* Input: pair of premise (element text span) and hypothesis (class description) (*n=48,000*) Output: 0 (neutral) or 1 (entailment)	gpt-5-mini
4	Hazard classification	Identifying hazard(s) addressed, if applicable		
5	Sector classification	Identifying targeted sector(s), if applicable		
6	Intentionality prediction	Predicting the adaptation intentionality score	**Intentionality predictor (SLM 3)** Fine-tuned *ClimateBERT* (regression) Input: element text span (*n=11,890*) Output: 0-1 intentionality score	gpt-5-mini (ordinal classification)

Table 1: Overview of tasks and models in the annotation process

the scores predicted by the regression SLM. Although this does not allow for a direct accuracy comparison, we can assess the extent to which the two approaches lead to similar categorical interpretations.

4 Evaluation methods

We evaluate the three SLMs (Table 1, column C) across three dimensions: annotation *accuracy*, annotation *quality*, and environmental *sustainability*. In each dimension, we benchmark the models against alternative (L)LMs and IAA metrics that are evaluated through an identical process. Below, we unpack the evaluation process of each dimension.

4.1 Annotation accuracy

To determine the accuracy of the models' annotations, we take a stratified hold-out sample from the original hand-labelled dataset (see section 2) and use this as the gold standard for evaluating the SLM- and LLM-generated annotations. We compute separate scores to represent accuracy per task (Table 1). Each score is also computed on annotations created by different human annotators: during manual annotation, a set of 287 chunks were annotated by at least two annotators - from which a total of 3,149 policy elements were extracted and coded. As the subset used for multi-coding was determined prior to the manual annotation process, it contains chunks from both the train and test sets. The multi-coded sample is used to compute benchmark F1 scores of human IAA that the SLMs should ideally approximate. We choose to report F1 scores rather than precision and recall separately, as F1 is symmetric and can therefore be used for both model accuracy and IAA.

For a comprehensive understanding of the accuracy of extracting the text sequences that represent a policy element, we introduce three metrics: *token-level pooled* F1 score (tok_{pool}), *weighted sequence-level pooled* F1 score (seq_{pool}), and *weighted sequence-level macro-averaged* F1 score ($\text{seq}_{\text{macro}}$). To determine whether two policy elements extracted from the same input chunk by different annotators refer to the same entity, we account for partial matches by using a semantic similarity threshold. More details about the metrics and the similarity criteria are provided in Appendix E.

4.2 Annotation quality

Human-labelled data has traditionally been used as the gold standard for evaluating the output of NLP systems. In complex tasks, where IAA scores are typically moderate, such a black-or-white approach can be questionable. To assess how well the annotation accuracy reflects the perceived quality of automatically generated annotations, we conduct a blind review with four experts. Each expert rates a set of 170-250 chunks (ensuring that they do not rate their own annotations), where two variants of the annotated, full chunk were shown. Each vari-

ant was annotated by either a human annotator, the SLM framework, or the benchmark LLM; the rater does not have information on what variant is created by which annotator. Each pair is rated on three metrics: *completeness*, *element type accuracy*, and *element relevance*. Per metric, the raters indicated whether they strongly or mildly preferred annotation A or B, or whether they found them to be of equal quality.

4.3 Environmental sustainability

To estimate the environmental impact of the LM inferences, we use the *EcoLogits* Python library (EcoLogits). At the time of writing, this library does not account for the impact of input tokens. Therefore, we estimate the weight of the input tokens to be equal to the ratio of costs per input token divided by the costs per output token, based on the *OpenAI* pricing (OpenAI). This results in an adjusted formula that replaces the number of output tokens used to compute environmental impact:

$$tokens_{total} = 1/6 * tokens_{input} + tokens_{output}$$

Using this adjusted variable gives a total estimation of the differences in impact across different LMs, prompt detail, and model reasoning settings. In addition to the benchmark *gpt-5-mini* model, we estimate the costs of a scenario in which the larger *gpt-5* model would be used to process the entire dataset. To estimate the impact of inference on classification and regression models, as these involve encoder-only models that are run locally, we use the *CodeCarbon* Python library (CodeCarbon).

5 Results

Here, we discuss the observed strengths and weaknesses of each specialist SLM in the framework separately.

5.1 Annotation model

Element extraction (task 1) - the accuracy scores of both the specialist annotation SLM and the benchmark LLM do not match human IAA in annotating adaptation policy overall, but the differences are relatively minor (Figure 1. The pooled, token-level agreement between the LLM annotator and the human test set even matches the IAA (Figure 1a). On all three scores, the benchmark LLM shows to create more accurate annotations than the specialist SLM when taking the human-labelled data as the gold standard. Overall, however, the LLM tends to over-annotate, as reflected in the number of extracted policy elements: on average, the benchmark LLM extracts more than 7 policy elements per input chunk, while the human annotators extract 4.4. The specialist SLM seems stricted in its inclusion criteria than the LLM, hence an average of 5.5 elements per chunk. As a result, the precision of the specialist SLM is generally higher, but the wide coverage of policy elements by the LLM leads to a substantial advantage on the recall metric and, therefore, a slightly higher F1 score, particularly on the pooled metrics (Figure 1a,b). The minor difference of seq_{macro} (0.42 vs 0.45) and the relatively large margin to the IAA on that metric (Figure 1c) emphasizes that both models tend to struggle with lower-density chunks (i.e., texts where only few or no policy elements are found based on the gold standard), which contribute more strongly to seq_{macro} than to the other scores.

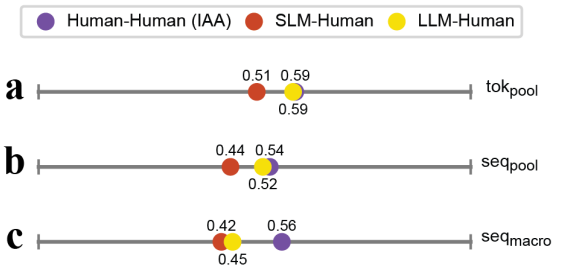


Figure 1: Accuracy and inter-annotator agreement on extracting adaptation policy elements (task 1). The metrics are *token-level pooled* F1 score (tok_{pool}), *weighted sequence-level pooled* F1 score (seq_{pool}), and *weighted sequence-level macro-averaged* F1 score (seq_{macro}). On each metric, the figure shows scores for IAA, the specialist SLM, and the benchmark LLM.

When analysing the *quality* of the generated annotations, the opposite trend emerges: although the benchmark LLM achieves higher accuracy than the specialist SLM, human raters prefer the specialist SLM based on perceived annotation quality. On the *element relevance* criterion, the SLM's annotations are rated significantly better than those of the LLM; the LLM is rated better on the *completeness* criterion, but this higher average rating is not statistically significant. Although human-created annotations are still rated significantly better than both models, the automatically generated ones are rated as being of equal (SLM - 39%; LLM - 26%) or better (SLM - 17%; LLM - 24%) quality than those of the human annotator (Figure 2). These substantial numbers underscore the need to step

beyond solely relying on accuracy metrics that take human-labelled annotations as the gold standard.

When it comes to environmental sustainability, a large difference emerges. Per set of 1,000 chunks, the specialist SLM consumes an estimated 0.3 MJ to process the chunks and return JSON objects of the text sequences and their element types (Appendix Figure 1). The gap in energy use to prompting general-purpose LLMs is substantial: when using the smallest variant of OpenAI's GPT-5 model (*gpt-5-nano*) with low reasoning effort, the consumed energy is estimated to be six-folded (1.6 MJ). Although this still sounds moderate, this model generated poor-quality annotations overall, where the returned text sequences were often not present in the input chunk or not adaptation-relevant. The energy use of the benchmark LLM (*gpt-5-mini* with low reasoning effort) averages 6.8 MJ for processing 1,000 chunks - 25 times as much as the SLM annotator. If one would increase reasoning effort [4] and/or use the larger *gpt-5* model to, potentially, improve annotation quality or accuracy, energy consumption would increase majorly (Appendix Figure 1). Processing all instances in the pre-filtered, adaptation-relevant dataset (70,000 chunks), consumed energy can approximate 10 GJ as opposed to less than 19 MJ when deploying the SLM annotator (Table 2). Furthermore, extending the prompt with a range of examples would increase input and reasoning tokens, extending the gap even further.

Besides energy use, the differences in water consumption footprint are equally substantial: while the specialist SLM needs approximately 7 liters of water to annotate the full dataset, this goes up to 161 (*gpt-5-mini*, low reasoning) and can even exceed 3,500 liters with the large *gpt-5* model (high reasoning). Figures of the average water consumption, global warming (GW) potential, and (primary) energy use can be found in Appendix F.

Type classification (task 2) - the annotation SLM not only extracted the text sequences: during the same inference, it assigned an element type (i.e., *Goal* or *Instrument*) to each identified element. On this task, the specialist SLM achieves accuracy comparable to the benchmark *ClimateBERT* model (Figure 3a). Both these models exceed the IAA scores, with F1 scores of 0.94 (*Goal*) and 0.83 (*Instrument*) as opposed to IAA F1 scores of 0.92 and 0.80 (Figure 3a). The other benchmark approach, classification with *gpt-5-mini* (where detailed descriptions of the *Goal* and *Instrument* classes were provided in the prompt) lead to poor results, with F1 scores below 0.6 for both element types (Figure 3a) - particularly due to the LLM being strongly skewed towards the *Instrument* class (Figure 3e). It is important to note that the sample sizes in the confusion matrices (Figure 3b-e) vary as the element type comparisons of the human-human and human-SLM pairs rely on the text spans that were identified by both annotators as extraction and type classification were combined, while the two benchmark models for task 2 (Table 1) took the same set of policy elements as input.

Although these scores may improve when extending the prompt with a varied set of examples for each element type (i.e., few-shot classification) and increasing reasoning, this would further increase energy consumption to disproportionate values for a relatively simple classification task, given that the benchmark *gpt-5-mini* model already has a substantial estimated primary energy use due to the large scale of the element-level dataset (Table 2).

5.2 Policy NLI classifier

The moderate agreement observed across human and computational annotators across the different classes assigned to policy elements (subtypes, hazards, and sectors) highlights the inherent complexity and ambiguity of in-depth adaptation policy annotation. Although the number of matches element pairs in the evaluation set is very low for some of the classes following the high class imbalance (Appendix B), benchmarking the class-level F1 scores of the policy NLI classifier against human IAA and the accuracy of an LLM classifier provides a general understanding of how well the classifier is capable of assigning the correct category to a given policy element.

Among the *Goal* subtypes, the NLI classifier approximates or even exceeds IAA: the NLI classifier and a human annotator mix up similar classes as two human annotators typically do (5a). The robustness of the scores of the *Instrument* subtypes is limited considering the sample size, but the NLI classifier, again, seems to approximate IAA and confuses similar classes as human annotators do (5b,c). Unpacking the most substantial mismatches (*n=94*) between the NLI classifier and the human-labelled standard highlights the capabilities of the model: in the majority (*Goal* - 62%; *Instrument* - 87%) of the evaluated mismatches, review re-

[4] Different reasoning settings for *gpt-5-nano* and *gpt-5-mini* were evaluated on the test set in the initial research phase. No substantial performance gains were observed.

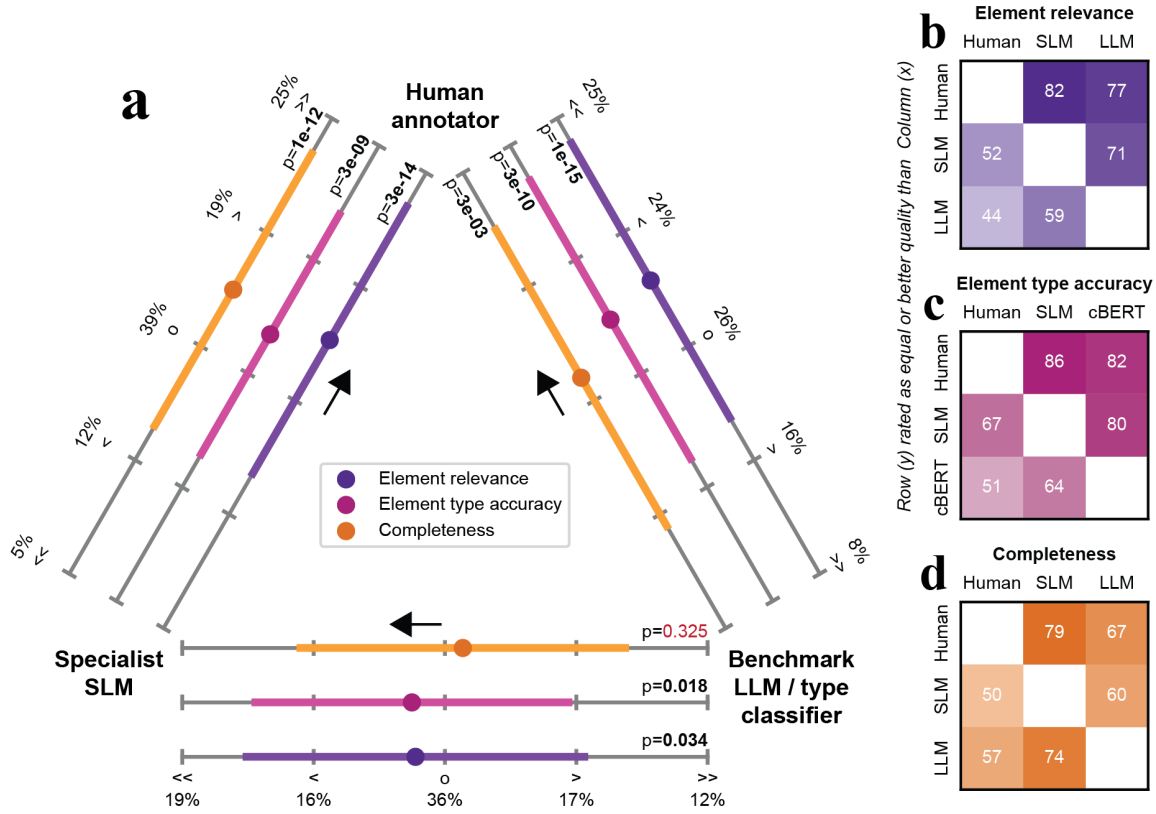


Figure 2: Overview of annotation ratings of a blind, comparative review. The triangle network (a) shows the mean and standard deviation of the 5-point scale ratings. The p-value to indicate whether the preference for a specific annotator (human, SLM, or LLM – indicated by the black arrows) is added to each scale. The percentage of raters selecting each option («, <, o, >, or ») is also indicated. The heatmaps (b, c, d) show the percentage of reviewers rating the row annotator (y axis) to be of equal or better quality compared to the column annotator (x axis), per criterion (element relevance, element type accuracy, and completeness).

vealed that the gold standard was incorrect (see Appendix H for more details).

On average, human annotators assigned a hazard category to only 14% and a sector category to 35% of the extracted policy elements. The number of instances in the datasets used to determine IAA and accuracy are, therefore, very limited. Among the few matched elements, the human annotators turned out to rarely agree on any explicit hazard being found at all, let alone on the specific category. Although relatively good recall is reported based on the NLI classifier's predicted hazards and sectors (Appendix Figure 6a,d), the model links a hazard and sector to a substantially larger share of the policy elements in the test set, particularly to the *environmental* and *meteorological (climate)* hazard classes, resulting in poor class precision (Appendix Figure 6a,d).

Following the lack of consensus among the annotators, the labelled hazard classes cannot be reliably interpreted as the gold standard. When mapping (dis)agreement between the NLI classifier and the benchmark *gpt-5-mini* classifier, we observe that the two models agree in most cases, except for the *meteorological (climate)* class, which the benchmark LLM assigns to a larger share of the policy elements to which the NLI classifier did not assign any hazard (Appendix Figure 6c). Although a reliable comparison between the NLI and LLM classifiers would require additional (blind) review, the excessive estimated environmental cost (Table 2), given the large number of policy elements in the global-coverage dataset, makes the *gpt-5-mini* model an unsustainable candidate for these classification tasks. Meanwhile, the estiamted consumption of the NLI model, despite requiring multiple inferences per policy element to cover all possible classes in premise-hypothesis pairs, remains negligible on the global-coverage dataset (Table 2).

5.3 Adaptation intentionality predictor

The regression model predicts the adaptation intentionality score with a MSE of 0.02 across the

Task	Model	Energy use (kWh)	Water cons. (L)	GW pot. ($kgCO_2eq$)	Prim. energy use (MJ)
Task 1 (element extraction)	Annotation model (SLM 1)	2	7	<1	19
	Benchmark LLM (gpt-5-mini, low R)	48	162	20	113
	Major LLM (gpt-5, high R)	980	3542	388	9590
Task 2 (type classification)	Annotation model (SLM 1)	N/A (merged in *Task 1*)			
	Benchmark cBERT classifier	0.05	0	0.01	
	Benchmark LLM (gpt-5-mini, low R)	48	173	25	549
Task 3-5 (subtype/hazard/ sector classification)	Policy NLI classifier (SLM 2)	0.5	0	0.1	
	Benchmark LLM (gpt-5-mini, low R)	198	712	103	2252
Task 6 (intentionality prediction)	Intentionality predictor (SLM 3)	0.04	0	0.01	
	Benchmark LLM (gpt-5-mini, low R)	45	162	23	512

Table 2: Estimated total environmental cost per model. For each SLM and benchmark LLM (reasoning setting indicated by *R*), the table shows the estimated consumption when processing the full dataset of adaptation-relevant chunks (task 1, *n=70,000*) and/or policy elements (tasks 2-6, *n=375,000*)

847 policy elements in the test set, while the score differences among human annotators are larger on average (*MSE=0.039*). Although partial matches in the IAA dataset were taken into account, likely reflecting slight differences in annotator interpretations and, consequently, intentionality scores, the SLM's substantial advantage over the IAA benchmark shows that it has learned from the underlying patterns in the training data to accurately predict adaptation intentionality. This capability is invaluable to determine whether policy addresses adaptation intentionally, implicitly, or indirectly.

When prompting *gpt-5-mini* to perform ordinal classification and mapping the classes onto the human-assigned (standardized) intentionality scores, we see that the LLM classifies most elements as *intentional* - even those that the human annotator assigned a relatively low score (Appendix Figure 7a). The predictions of the SLM and *gpt-5-mini* are more aligned, particularly showing a clearer distinction between indirect and implicit adaptation on the SLM's intentionality scale (Appendix Figure 7b). Again, the difference in environmental cost between the specialist SLM and benchmark LLM are large (Table 2) - predicting the intentionality scores with the *ClimateBERT* regression model leads to very minimal energy (<0.01 kWh) and no water consumption.

6 Discussion

Now that even smaller LMs have gotten so advanced, relying solely on accuracy metrics based on human-created gold standards has become questionable. Our results reveal how challenging achieving high IAA is on a complex, free-text annotation tasks - despite extensive training and iterative fine-tuning of the code book. Considering this moderate overall agreement, a computational annotator that perfectly matches the human gold standard, comprising annotations from multiple annotators, is not a feasible target here. Nonetheless, we show that LMs reach annotation accuracy that approximates IAA and produce annotations that are most often equal to - or even exceed - human-created annotations on perceived quality. Although overall accuracy scores of using a large, general purpose LLM for extracting policy element candidates come close to human IAA compared to a fine-tuned SLM, the consistency and stricter inclusion criteria

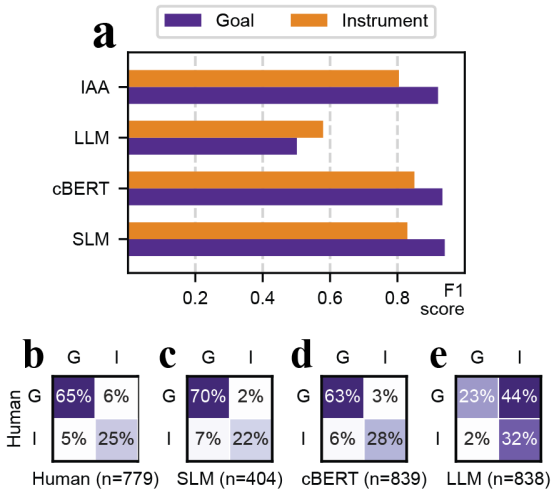


Figure 3: Accuracy and IAA on predicting element types (task 2). The bar chart (a) shows the F1 score of each element type (Goal and Instrument). The confusion matrices (b-e) show the percentages of agreement and confusion across human annotators (a) and between a model and human annotator (c-e).

of the specialist model makes its annotations being perceived as better quality overall. These observations underscore the importance of considering broad evaluation criteria that go beyond traditional accuracy metrics.

In addition to its perceived quality advantage, using a specialist SLM for this relatively complex annotation task brings a substantial advantage in terms of environmental cost: according to the *EcoLogits* estimations, processing all 70,000 chunks from a global-coverage evidence base with the specialist annotation model (SLM 1) as opposed to using *gpt-5-mini* saves 43 kWh - equivalent to approximately one week of electricity use by an average Dutch household (Centraal Bureau voor de Statistiek, 2025) - and more than 150 liters of water. Based on our accuracy and quality evaluations, the relatively small *gpt-5-mini* is not sufficiently advantageous over the specialist SLM to justify this cost. Opting for the larger and more advanced *gpt-5* model, however, leads to excessive and unsustainable consumption that could power a multitude of Dutch households for a full month and cover several days of water use.

Environmental cost reaches even higher levels when considering the use of LLMs for detailed classification tasks. Even when using a relatively small general-purpose LLM like *gpt-5-mini*, energy and water use and global warming potential becomes extremely disproportionate. All fine-tuning processes add only 1 kWh energy use combined (see Appendix D), further emphasizing the major gap in environmental cost between specialist SLMs and deploying general-purpose LLMs. This emphasizes that researchers should be aware and critical of the implications of using excessively large LMs in their research - particularly for structured tasks and in domains where evidence is even more voluminous (e.g., climate change mitigation). Frequent updating of the evidence base, which is inevitable given the continuous development of new climate policy documents and a field that moves towards living evidence platforms and databases that are updated frequently and automatically (Ford et al., 2025), would mean that environmental cost expands even further with every update cycle.

Taking into account the core intention of large-scale annotation of adaptation policies, where addressing climate change and sustainability logically plays a major role, we conclude that bigger is not better in the context of climate (adaptation) policy annotation. Considering that our specialist annotation model is only marginally behind human annotators in terms of accuracy and quality, we confidently suggest the use of fine-tuned SLMs as a sustainable, consistent, transparent, and accurate opportunity for upscaling climate policy annotation.

For related applications, it is important to note that fine-tuning an SLM to become an expert annotator requires sufficient hand-labelled examples. Establishing such a dataset may not always be feasible. In such cases, an alternative approach could be to use an advanced LLM to create a set of training examples - of which a representative sample should be manually checked - and use this to fine-tune an SLM. Depending on the size of the full dataset that needs to be annotated, this approach can majorly decrease environmental and financial costs.

Leveraging our proposed framework of SLMs as expert policy annotators moves beyond the constraints of manual expert annotation and enables processing adaptation policy data at large scale. Through this, it is possible to establish a global-coverage, structured evidence base that allows for systematically tracking adaptation policy progress across the globe, invaluable for global assessment such as the UNFCCC global stocktake (2028) and IPCC AR7 (2028-2029).

Limitations

There are some limitations to our study that are important to consider. As we relied on matched policy elements (i.e., text sequences identified by two annotators) to compute subtype, hazard, and sector IAA, while the classes were imbalanced (see Appendix B), the data used for testing the NLI model was limited. For a more robust estimation of IAA and model accuracy, particularly on the hazard classes, it is recommended to further expand the test dataset by double-coding a pre-defined set of policy elements.

As the parameters of the *OpenAI* LLMs are not publicly available, the environmental cost of inference with these models was based on estimations. The actual environmental cost may, therefore, somewhat deviate from the numbers presented in our paper. Selecting a different LLM that is not a GPT-5 model, may also influence the results. Furthermore, we have used a rough estimation of the cost of fine-tuning the SLMs in our comparison, as this was not measured during the processes. However, even with a large deviation, the environmental cost is still negligible considering the evident, disproportionate consumption of LLMs.

Finally, we have not accounted for uncertainty in the generated annotations. Uncertainty quantification during manual annotation and in the computationally generated annotations (e.g., by specifying confidence scores), however, could bring an interesting perspective on certain cross-annotator (mis)alignment patterns.

Acknowledgements

JB and RB acknowledge funding support from the Dutch Research Council (NWO VI.Vidi.211.132). Furthermore, we express our gratitude to the team of annotators - Robin Blersch, Denatalie Chrisdameria, Demi Ursem, and Suze Zander - and thank them for their important contributions to establishing the hand-labelled dataset. Lastly, we thank the anonymous reviewers for their detailed review on a previous version of this manuscript.

References

Jetske Bonenkamp, Robbert Biesbroek, and Ioannis N. Athanasiadis. 2025. Transforming adaptation tracking: Benchmarking Transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In *Proceedings of The 2nd Workshop of Natural Language Processing meets Climate Change*.

Centraal Bureau voor de Statistiek. 2025. Energieverbruik particuliere woningen; woningtype en regio's. Last Modified: 2025-09-17T02:00:00+02:00.

Climate Policy Radar. Climate Policy Radar | AI for climate law and policy research.

CodeCarbon. CodeCarbon.

EcoLogits. EcoLogits | Estimate and track the environmental footprint of GenAI models at inference.

James D. Ford, Robbert Biesbroek, Lea Berrang Ford, Felix Creutzig, Neal Haddaway, Sherilee Harper, Jan C. Minx, Mark New, Anne J. Sietsma, Carol Zavaleta-Cortijo, and Max Callaghan. 2025. Recommendations for producing knowledge syntheses to inform climate change assessments. *Nature Climate Change*, 15(7):698–708.

Anandita Garg, Uma Gaba, Deepan Muthirayan, and Anish Roy Chowdhury. 2025. Emissions and Performance Trade-off Between Small and Large Language Models. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 515–520.

Daniel Henstra. 2016. The tools of climate adaptation policy: analysing instruments and instrument selection. *Climate Policy*, 16(4):496–521. _eprint: https://doi.org/10.1080/14693062.2015.1015946.

IPCC. 2022. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1 edition. Cambridge University Press.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. Large Language Models as Span Annotators. *arXiv preprint*. ArXiv:2504.08697 [cs].

OpenAI. API Pricing.

Ben Orlove. 2022. The Concept of Adaptation. *Annual Review of Environment and Resources*, 47(1):535–581.

Ana Petek, Nikola Baketa, Anka Kekez, Marko Kovačić, Mario Munta, Krešimir Petković, Marjeta Šinko, and Borna Zgurić. 2021. Unboxing the vague notion of policy goals: Comparison of Croatian public policies. *European Policy Analysis*, 7(2):451–469.

Nghiem Thanh Pham, Tung Kieu, Duc-Manh Nguyen, Son Ha Xuan, Nghia Duong-Trung, and Danh Le-Phuoc. 2025. SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Pilar Bueno Rubial, Anand Patwardhan, María Luz Falivene Fernández, Joel González, Victoria Laguzzi, Susana Zazzarini, and Carolina Passet. 2023. UAE Framework for Global Climate Resilience: key outcomes of COP28 and pathways towards COP29 and COP30.

Isabel Segura-Bedmar, Paloma Martinez, and Maria Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Seventh International Workshop on Semantic Evaluation*, volume 2, Georgia.

Sebastian Sewerin, Lynn H. Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. Towards understanding policy design through text-as-data approaches: The policy design annotations (POLIANNA) dataset. *Scientific Data*, 10(1):896.

Anne J. Sietsma, James D. Ford, and Jan C. Minx. 2024. The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, 14(1):31–39.

UNDRR. 2020. Hazard definition and classification review: Technical report. Technical report.

UNFCCC. 2025. Final list of potential indicators, UAE-Belém work programme on indicators.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. TnT-LLM: Text Mining at Scale with Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5836–5847, New York, NY, USA. Association for Computing Machinery.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):145:1–145:87.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text. *arXiv preprint*. ArXiv:2110.12010 [cs].

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, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs.CL].

Appendices

A Class descriptions

Subtypes

Each policy element can be of type *Goal* or *Instrument*. Within these element types, subtypes are defined, which are described in detail in Appendix Table 1. The *Goal* subtypes are based on the policy categories that Petek et al. (2021) defined. The *Instrument* subtypes combine elements of the resources of Henstra (2016) and the taxonomy of POLIANNA (Sewerin et al., 2023).

Hazards and sectors

Policy elements can be assigned a climatic hazard or sector, where applicable. These hazard and sector classes are provided in Appendix Table 2. The typologies are based on the general hazard clusters of the UNDRR (2020) and the impacts of climate change on ecosystems and human systems from IPCC's Sixth Assessment Report (IPCC, 2022).

B Class distribution

An overview of the distribution of classes is provided in Table 3 (subtypes) and Table 4.

Goal (69%)		**Instrument** (31%)	
INS	3138 (39%)	ORG	2436 (68%)
PRO	2831 (35%)	KNO	598 (17%)
SEC	1987 (24%)	TEC	213 (6%)
VAL	107 (1%)	REG	153 (4%)
EVA	78 (1%)	PUB	103 (3%)
		FIN	100 (3%)

Appendix Table 3: Distribution of subtype classes in the hand-labelled dataset

Hazard		**Sector**	
None	10280 (86%)	*None*	7755 (65%)
MET_c	747 (6%)	NAT	1061 (9%)
GEO	411 (3%)	FOO	992 (8%)
BIO	167 (1%)	WAT	876 (7%)
MET_w	161 (1%)	BUI	764 (6%)
ENV	124 (1%)	HEA	442 (4%)

Appendix Table 4: Distribution of hazard and sector classes in the hand-labelled dataset

Type	Subcategory	Definition
Goal		*Quantitative or qualitative objective that a party aims to achieve, with the overarching aim of adapting to climate change. Involves a desired/aspired outcome or intended state/action*
	SEC: Sector-oriented	Goals directly aimed at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region (accessibility and security of goods and services; well-being and prevention; development and mobility)
	PRO: Process-oriented	Goals that focus on the organizational aspect of the regarding policy, only indirectly intended to address an issue (cooperation; capacity-building; strategic planning)
	EVA: Evaluation-oriented	Goals that involve meeting a set evaluation criterion in a given sector (quality; sustainability; effectiveness)
	INS: Instrument-oriented	Goals that refer to the introduction, change, or continuation of a policy instrument as a policy goal (regulation; investment; (re)organization; knowledge and awareness)
	VAL: Value-oriented	Goals that aim at ensuring or advancing core constitutional values of democracies or political fundamentals (justice and law; environmental protection; rights and equality)
Instrument		*The means to achieve the defined targets, typically in the form of a tool, mechanism, or strategy. Specifies how a goal can be achieved and can be written in present or past tense*
	REG: Regulation and legislation	The use of the state's legal authority to order, forbid, or permit certain behaviour
	FIN: Financial incentive	Taxes, subsidies, or tradable allowances to stimulate behaviour or activities
	ORG: Organization and cooperation	The use of governmental resources through programmes, services, plans, and/or (voluntary) agreements
	PUB: Public investment	Strategic investments and the management thereof, done by public bodies
	KNO: Knowledge and information	Research, education, training, and knowledge sharing, focused on knowledge generation and increasing awareness
	TEC: Technology and innovation	Creation, advancement, and commercialization of new technologies and innovative systems or incentives

Appendix Table 1: Overview of the subtype typology for annotating adaptation policy.

Type	Subcategory	Definition
Hazard		*A policy element gets assigned a hazard only if one can clearly identify what type of hazard it addresses by the text sequence itself or the context (i.e., the surrounding text in the same chunk)*
	MET_w: Meteorological (weather)	Addresses problems rising from extreme weather events produced by the Earth's state and behaviour, occurring during a short period (minutes to days) and/or on small scale (storms, extreme temperatures, heavy rain/snowfall, heatwaves, short-term droughts)
	MET_c: Meteorological (climate)	Addresses problems rising from changes in weather patterns and/or climatic conditions produced by the Earth's state and behaviour. These typically span a longer period and/or occur at large scale (shifting precipitation patterns, rising average temperatures, long-term droughts, wildfires)
	GEO: Geohydrological	Addresses problems rising from mass transfer of water or solid earth, caused by the interaction of atmospheric conditions with land and ocean systems (floods, landslides, glacial melting, coastal erosion)
	ENV: Environmental	Addresses problems rising from degradation of ecosystem services and natural systems, posing threat to water, land, air, biodiversity, as well as crucial earth processes (deforestation, soil degradation, air pollution, desertification)
	BIO: Biological	Addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna (vector-borne diseases, invasive species, animal epidemics)
Sector		*A sector is only assigned to a policy element if the link is explicitly stated or can be identified through the surrounding context*
	FOO: Food production	Addresses any of the following sub-sectors: agriculture/crop production; animal and livestock health and productivity; fishery yields and aquaculture production
	WAT: Water resources	Addresses any of the following sub-sectors: availability, accessibility, quality, of water; management of water resources
	HEA: Health and wellbeing	Addresses any of the following sub-sectors: air quality; heat; mental health; malnutrition; infectious diseases
	BUI: Built environment	Addresses any of the following sub-sectors: infrastructure, cities, settlements, communities, tourism, or economic sectors
	NAT: Natural environment	Addresses any of the following sub-sectors: terrestrial, freshwater, or ocean ecosystems; biodiversity; phenology

Appendix Table 2: Overview of the hazard and sector typology for annotating adaptation policy.

C Examples of premise-hypothesis pairs

A set of examples of the premise-hypothesis pairs, used to fine-tune the policy NLI agent, is provided in Appendix Table 5.

D Fine-tuning details

The annotation model (SLM 1) was fully fine-tuned for 3 epochs using the *AdamW* optimizer with an effective batch size of 32 (per-device batch size - 4; gradient accumulation steps - 8) and a learning rate of 5*e-5*. We used mixed-precision (FP16) training and gradient checkpointing to reduce required memory. The policy NLI agent and adaptation intentionality predictor were fine-tuned with similar settings, except for the (per-device) batch size (16). Finally, the *ClimateBERT* model fine-tuned as the benchmark model for element type classification was fine-tuned with a lower learning rate (5e-6).

Fine-tuning was performed on a single *NVIDIA Tesla V100-SXM2-16GB* GPU, which has a rated maximum power consumption of 300 W. When operating at maximum power consumption continuously, the estimated power consumption per hour of fine-tuning equals 0.30 kWh. Considering that the fine-tuning processes ranged between 20 minutes and 2 hours for the SLMs in our framework, we roughly estimate the total power consumption of fine-tuning all three models to approximate 1 kWh.

Premise (policy element)	Class	Hypothesis (class description)	Label
promoting climate-resilient agricultural practices	Food production (sector)	the element addresses agriculture/crop production, animal and livestock health and productivity, fishery	1 (entailment)
At least 3 complete sets of specialised hazards maps & detail analysis in place	Sector-oriented (Goal subtype)	the element aims at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region	0 (neutral)
provision of financial assistance to forest communities affected by climate change	Instrument-oriented (Goal subtype)	the element involves the introduction, change, or continuation of a policy instrument as a policy goal	1 (entailment)
Plan for Implementing Climate Change Science	Biological (hazard)	the element addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna	0 (neutral)

Appendix Table 5: Examples of premise-hypothesis pairs used for fine-tuning a NLI-inspired classifier.

E Methods per annotation task

Here, we provide an overview of the detailed methodology for each of the tasks in the annotation process.

Task 1 (element extraction) – evaluating the accuracy of extracting the text sequences that represent a policy element comes with several challenges and can be approached from several angles and scales. To get a comprehensive understanding of the annotators' sequence extraction accuracy, we introduce three metrics: token-level pooled F1 score (tok_{pool}), weighted sequence-level pooled F1 score (seq_{pool}), and weighted sequence-level macro-averaged F1 score (seq_{macro}).

Since the task involves free-text annotation, using only exact-boundary matches may lead to correct, partial matches being missed (Segura-Bedmar et al., 2013). Furthermore, different interpretations can lead to annotators splitting sequences differently, but a crucial part of establishing the evidence base is that the right parts of the text are extracted. The tok_{pool} metric, therefore, zooms out to the full set of policy elements per annotator, from which it creates a list comprising all tokens that are not a stop word (e.g., "the" or "at"). Following, the lists of tokens of annotator A and B are compared by counting tokens as true positive (token from list B that also occurs in list A), false positive (token from list B that does not occur in list A), or false negative (token from list A that does not occur in list B) and computing the tok_{pool} F1 score.

While tok_{pool} helps to understand how well the annotators extract the right information, the sequence scores (seq_{pool} and seq_{macro}) measure the accuracy of the extracted policy elements as units. Considering that a boundary mismatch does not always mean that two sequences do not refer to the same entity (e.g., "the new national climate strategy of 2050" versus "national climate strategy of 2050"), we assess semantic similarity of two sequences (of different annotator on the same input chunk) if they have overlapping tokens. If the semantic similarity score exceeds 0.XX, they are considered to be referring to the same policy element, which allows for counting each policy element as a true positive, false positive, or false negative.

For the pooled variant (seq_{pool}), the F1 score is based on the total set of policy elements per annotator, indicating to what extent the extracted policy elements overlap. To this score, dense chunks (i.e., text blocks in which many policy elements are found) contribute more strongly compared to chunks in which only a few (or no) policy elements are identified. In contrast, the macro-averaged

score ($\text{seq}_{\text{macro}}$) takes the average of the F1 score per chunk, enabling an understanding of the extent to which the chunks are processed equally, and, e.g., correctly extracting zero elements from a chunk in which no relevant adaptation policy is found is rewarded. Here, the annotation accuracy of each chunk contributes to the score equally, regardless of the number of policy elements extracted.

The seq_{pool} and $\text{seq}_{\text{macro}}$ scores both use weights, which are determined by the adaptation intentionality scores of the policy elements. An element that intentionally addresses adaptation (e.g., building flood defence mechanisms) is considered to be more crucial not to miss out on compared to more indirect adaptation measures (e.g., climate awareness programmes). The total counts used as true positives, false positives, and false negatives are, therefore, the sums of the elements' intentionality scores, resulting in weighted F1 scores. The sum of true positives and false negatives in case of no overlapping policy elements in a single chunk, when computing seq_{pool}, forms an exception. In such cases, the F1 score will always equal zero, and there is no difference in score between, e.g., a single false negative with a low intentionality score as opposed to, e.g., five intentional policy elements being missed. We therefore introduce a custom formula to substitute the weighted sum of true positives to compute seq_{pool}:

$$TP_{substitute} = 1/(1+FN_{weighted}), \quad if\ TP = 0$$

$$FN_{substitute} = 1/(1+FP_{weighted}), \quad if\ FN = 0$$

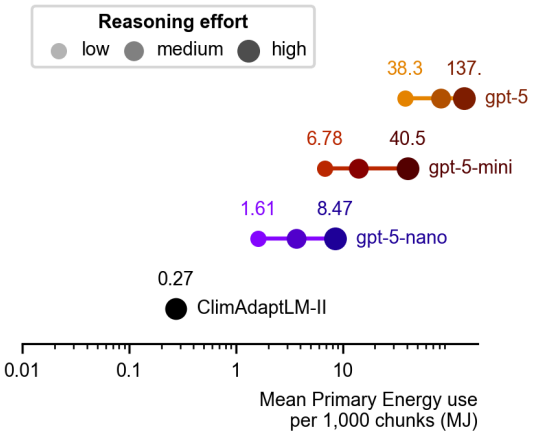
To illustrate, when annotator A did not identify any policy elements, and annotator B identified one element with an intentionality score of 0.4, the true positive substitute would equal 0.71, leading to a recall of 0.64. In case annotator B identified six elements with an average intentionality score of 0.75, however, the substitute equals 0.21, hence a recall of 0.04 – penalizing the missing annotation overlap more strongly. If both annotators identified no elements in a given chunk, we take an F1 score of 1 for that chunk.

Task 2-5 ((sub)type, hazard, and sector classification) – based on the matched policy elements arising from evaluation of the first task, the F1 scores of predicting the type (i.e., *Goal* or *Instrument*), subtype (e.g., process-oriented goal), hazards, and sectors are computed. F1 scores are reported per class.

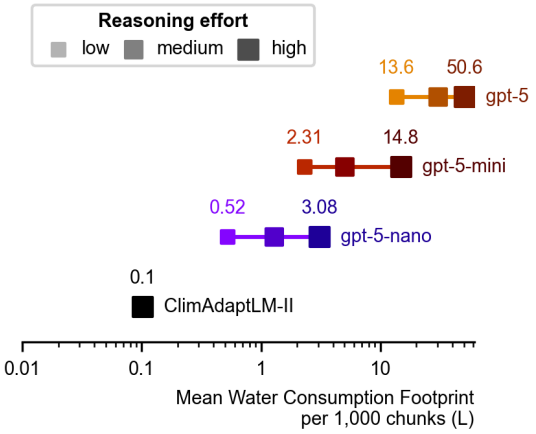
Task 6 (intentionality prediction) – to determine the direct accuracy of the regression model, we compute the Mean Squared Error (MSE). Additionally, the intentionality class labels assigned by the benchmark LLM (*gpt-5-mini*) are aligned to the intentionality scores of the SLM and human annotator.

F Environmental impact

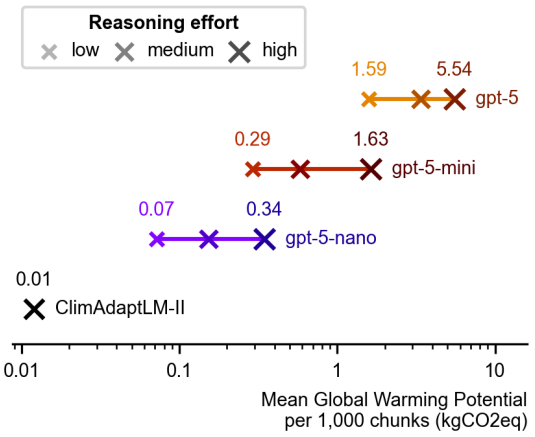
See Appendix Figure 1, Appendix Figure 2, Appendix Figure 3, and Appendix Figure 4.



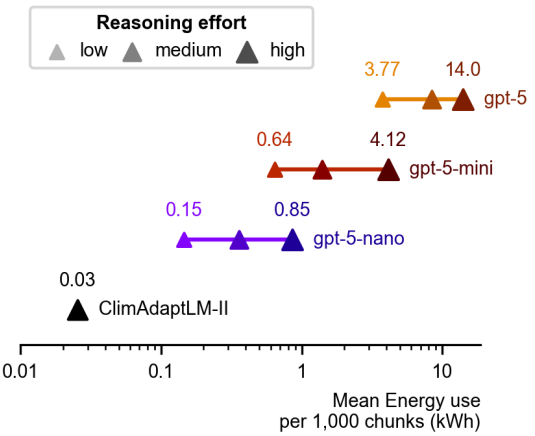
Appendix Figure 1: Estimated mean primary energy use per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



Appendix Figure 2: Estimated mean water consumption per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



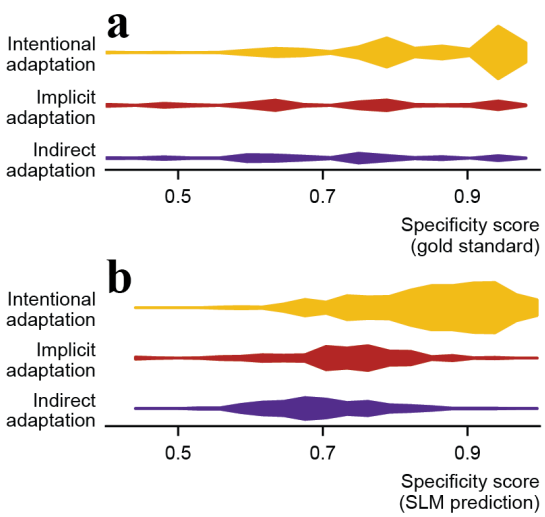
Appendix Figure 3: Estimated mean global warming potential per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



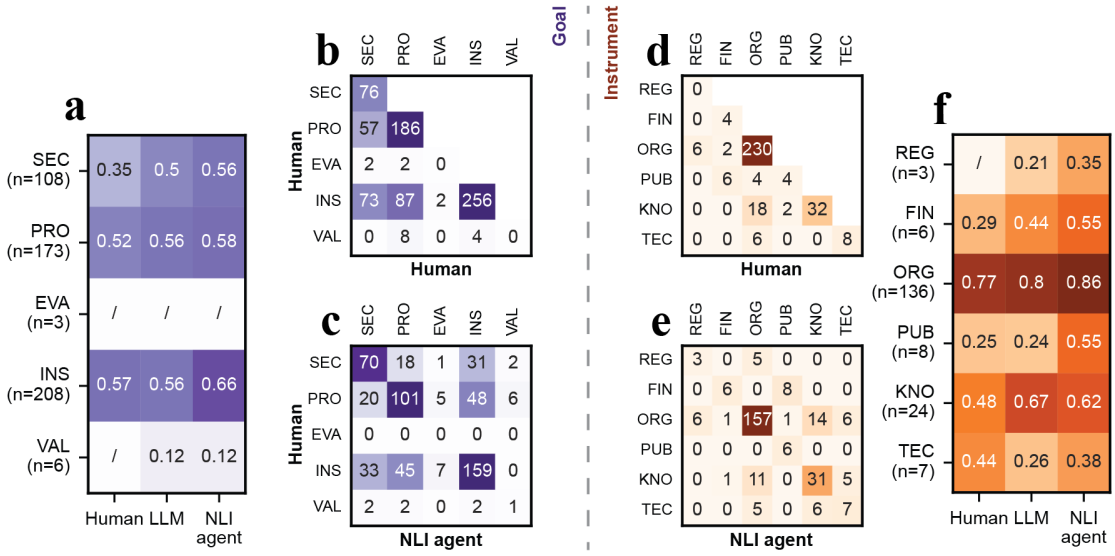
Appendix Figure 4: Estimated mean energy use per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.

G Class agreement

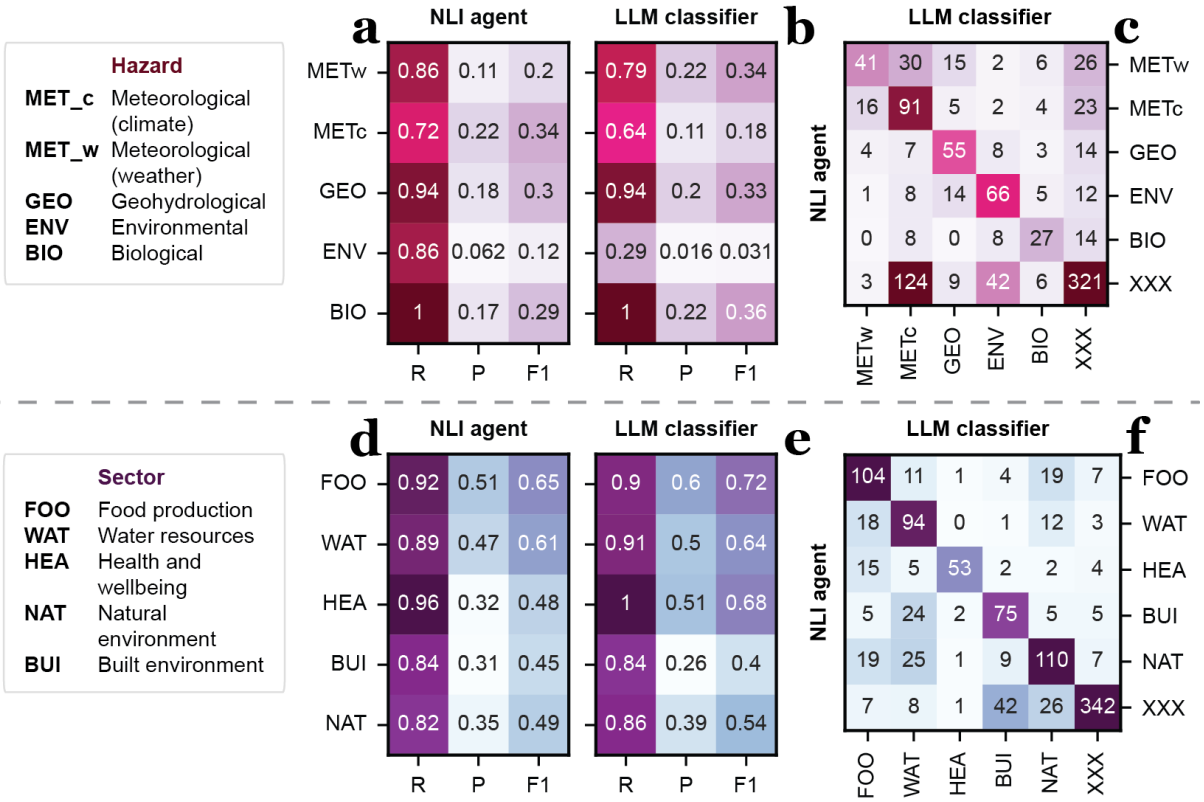
See Appendix Figure 5, Appendix Figure 6, and Appendix Figure 7.



Appendix Figure 7: Counts of intentionality class predictions by an LLM classifier per intentionality score interval. The filled areas show the total number of policy elements classified as the intentionality class (y axis) of each specificity score interval, determined by splitting the scores into 20 equal bins.



Appendix Figure 5: Accuracy and inter-annotator agreement of predicting subtypes (task 3). Subfigures (a) and (f) show the F1 scores per annotator (x axis), per class (y axis). The number of elements of that class in the test set is printed below the class labels (y axis). The confusion matrices show the confusion of two human annotators (b and d) and the NLI agent versus a human annotator (c and e) on the Goal (left) and Instrument (right) subtypes.

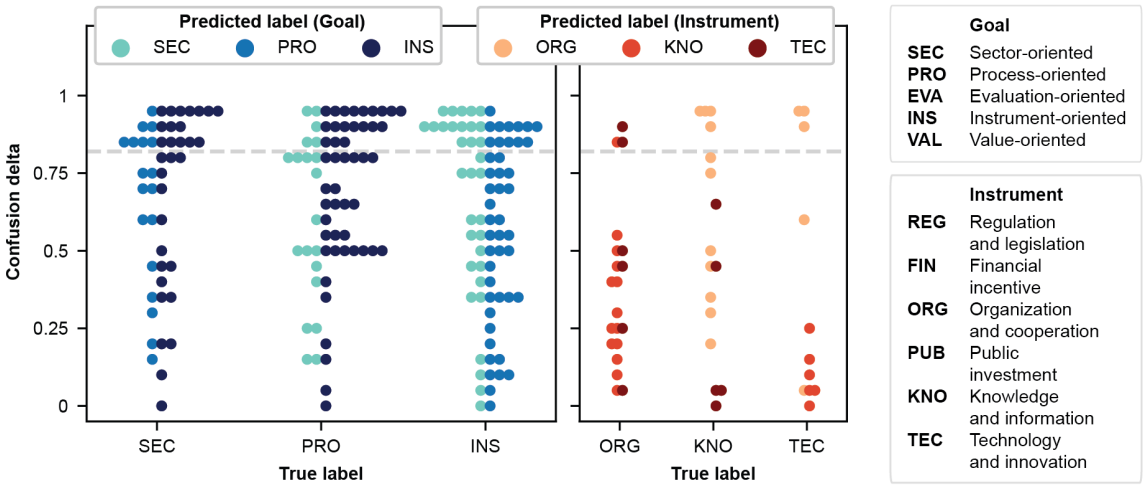


Appendix Figure 6: Accuracy of the policy NLI classifier and benchmark LLM on hazard and sector classes. Subfigures (a), (b), (d), and (e) show the recall, precision, and F1 score of the NLI agent (a and d) and benchmark LLM against the human-labelled gold standard on the hazard (a and b) and sector (d and e) classes. The confusion matrices in (c) and (f) show the agreement between the NLI agent and the benchmark LLM on the hazard (c) and sector (f) classes.

H Subtype mismatch analysis

To understand disagreement between annotators on class labels, we assessed substantial mismatches between the gold standard and predicted subtype (task 3) in detail. The selected mismatches are those where the confusion delta (i.e., the difference in probability of the gold standard and predicted label based on the NLI classifier) exceeds 0.8 (above the dashed lines in 8). For this process, we have prepared a dataset in which the two labels (gold standard are predicted) are randomly shuffled for review to account for possible bias. The review dataset consists of 79 annotated goals and 15 annotated instruments. On each policy element, we have indicated whether label A is correct, label B is correct, or both apply, revealing the extent to which deviations from the gold standard are incorrect, human errors, or a result of ambiguity.

Among the 79 reviewed *Goal* elements, both classes were considered applicable in 10 cases, as opposed to 6 cases where none of the two labels were relevant according to the reviewer. Out of the remaining reviewed instances, the label given by the specialist SLM was chosen 39 times (62%) while the human annotator's assigned label was considered correct 24 times (38%). Within the set of Instrument mismatches, the difference is even larger: only 2 of the human-assigned classes were rated as the correct label as opposed to 13 of the SLM-assigned classes (87%).



Appendix Figure 8: Subtype mismatches between gold standard (x axis) and predicted label. Each bubble represents a label mismatch, where the y location indicates the difference in model confidence on the predicted and true label.

I Model prompts

I.1 Element extraction (task 1)

```
# Instruction

You are an expert in extracting climate
policy elements from text.  Your task
is to retrieve text sequences that
represent a policy element that relates
to climate change from the input text.
Make sure to return JSON objects that
contain verbatim substrings and do not
correct, modify or paraphrase the text
in any way.  Do not include punctuation
at the end of the substring.  Assign
the most suitable policy type to each
object.

## Policy element

A policy element is a sequence of words
that refers to a goal, instrument, or
output that can be seen as a means to
contribute to an overarching objective,
in this case any element related to the
**climate change** problem.

It might be that the provided text does
not contain any policy elements.  If so,
return an empty array.

### Inclusion criteria

1.  Direct reference to climate impacts:
there is an explicit mention of risks,
hazards, or vulnerability linked to
climate change (e.g., early warning
systems for floods).
2.  Focus on building resilience or
vulnerability reduction:  the element
focuses on improving the ability of
communities or (eco)systems to limit
the impact (proactive) or recover
(reactive) from expected climate impacts
(e.g., objectives to reduce drought
vulnerability in dry areas)
3.  Indirect link:  the element does
not include a clear link to intentional
adaptation or mitigation - but is known
to indirectly contribute to, e.g.,
adaptive capacity, increased resilience,
and/or reduced vulnerability (e.g.,
sustainable land use, climate research
programs, climate subsidies for farmers)
4.  Alignment with broader objectives:
the element is mentioned to contribute
to or support an overarching (e.g.,
international) objective through a
specific goal or means (note:  not the
mention of the overarching objective
itself)
5.  Introduction of a new climate policy
or inclusion of climate change in an
existing policy.  This also includes,
e.g., programmes and projects targeted
at climate change
```

```
### Exclusion criteria

1. Unrelated or general development policy: the element describes actions that are not specifically linked to climate risks (e.g., building schools or reducing poverty). If it does focus on general climate change (e.g., prioritizing climate change in school programs) or is listed as a subitem of climate change policy (e.g., reducing poverty to increase resilience), however, it should be included (see item 3 of the inclusion criteria)
2. Disaster risk reduction elements that focus on short-term recovery or non-climate-related risks
3. Support to other countries: the retrieved policy elements should have a **national** focus
4. Overview of the contents of a document or policy: the element describes what another policy (document) consists of or targets, without providing a specific goal/instrument/output
5. Mention of the current document (title): if the text simply repeats the title or type of the current document, it should not be marked as a policy element

## Policy domain

Each climate policy element is assigned one of the following climate policy domains:

1. "Adaptation": refers to the process of adjusting natural or human systems in response to actual or expected climatic changes and their effects, in order to moderate potential harm, exploit beneficial opportunities, or cope with unavoidable consequences
2. "Mitigation" : refers to efforts to reduce or prevent the emission of greenhouse gases (GHGs) into the atmosphere, or to enhance the removal of these gases through natural or technological sinks. The goal of mitigation is to limit the magnitude or rate of long-term climate change itself, addressing its causes rather than its consequences
3. "General": these elements may target both adaptation and mitigation or be overarching elements (e.g., climate awareness)

## JSON schema

{ "type": "array", "items": { "type": "object", "properties": { "policy_element": { "type": "string" }, "policy_domain": { "type": "string", "enum": ["Adaptation", "Mitigation", "General"] } }, "required": ["policy_element", "policy_domain"] } }
```

I.2 Element type classification (task 2)

```
# Instruction

Your task is to classify adaptation policy elements as "Goal" or "Instrument". Only return "Goal" or "Instrument"; do not add any other text. Considering the following class definitions:

## Goal

*Description**: Quantitative or qualitative objective that a party aims to achieve, with the overarching aim of adapting to climate change. Involves a desired/aspired outcome or intended state/action

*Semantic indicators**: Use of ambitious and/or future-oriented language (‘to achieve’, ‘aiming to’, ‘target’) or terms typically used for goals (e.g., ‘increase’, ‘ensure’, ‘promote’, ‘support’)

## Instrument

*Description**: The means to achieve the defined targets, typically in the form of a tool, mechanism, or strategy. Specifies how a goal can be achieved and can be written in present or past tense
```

*Semantic indicators**: Mention of a specific tool, action, or intervention (e.g., 'regulation', 'program') or tangible action (e.g., 'build dykes')

I.3 Subtype classification (task 3)

Goal subtypes

Your task is to classify a climate change adaptation policy goal.

Pick the category (out of the 5 options) that best fits the provided goal. Return the three-letter code of the category without generating any other text.

Categories:
*SEC** - Sector-oriented goal.
This category applies when the policy element aims at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region. Related themes can be: accessibility and security of goods and services; well-being and prevention; development and mobility.

*PRO** - Process-oriented goal.
This category applies when the policy element involves the organisational aspect of the policy, a process indirectly intended to address an issue. Related themes can be: cooperation; capacity-building; strategic planning.

*EVA** - Evaluation-oriented goal.
This category applies when the policy element involves meeting a set evaluation criterion in a given sector. Related themes can be: quality; sustainability; effectiveness.

*INS** - Instrument-oriented goal.
This category applies when the policy element involves the introduction, change, or continuation of a policy instrument as a policy goal. Related themes can be: regulation; investment; (re)organisation; knowledge and awareness.

*VAL** - Value-oriented goal.
This category applies when the policy element aims at ensuring or advancing core constitutional values of democracies or political fundamentals. Related themes can be: justice and law; environmental protection; rights and equality.

Instrument subtypes

Your task is to classify a climate change adaptation policy instrument.

Pick the category (out of the 6 options) that best fits the provided instrument. Return the three-letter code of the category without generating any other text.

Categories:
*REG** - Regulation and legislation instrument.
This category applies when the policy element uses the state's legal authority to order, forbid, or permit certain behaviour.

*FIN** - Financial incentive instrument.
This category applies when the policy element uses taxes, subsidies, or tradable allowances to stimulate behaviour or activities.

*ORG** - Organisation and cooperation instrument.
This category applies when the policy element involves the use of governmental resources through programmes, services, plans, and/or (voluntary) agreements.

*PUB** - Public investment instrument.
This category applies when the policy element involves a strategic investment and/or the management thereof, done by public bodies.

*KNO** - Knowledge and information instrument.
This category applies when the policy element involves research, education,

training, and/or knowledge sharing, focusing on knowledge generation and/or increasing awareness.

*TEC** - Technology and innovation instrument.
This category applies when the policy element involves creation, advancement, and/or commercialization of new technologies or innovative systems or incentives.

I.4 Hazard classification (task 4)

Your task is to identify the hazard(s) addressed by a climate change adaptation policy element.

Pick 0, 1, or multiple categories (out of the 5 options) that fit the provided element.
Only select a category when the policy element **explicitly** addresses the hazard.
Return an array of the three-letter codes of the categories, without generating any other text.
If no hazard category fits the adaptation policy element, return an empty array ([]).

Categories:
*MET_w** - Meteorological (weather) hazard.
This category applies when the policy element addresses problems rising from extreme weather events produced by the Earth's state and behaviour, occurring during a short period (minutes to days) and/or on small scale. Related themes can be: storms, extreme temperatures, heavy rain/snowfall, heatwaves, short-term droughts.

*MET_c** - Meteorological (climate) hazard.
This category applies when the policy element addresses problems rising from changes in weather patterns and/or climatic conditions produced by the Earth's state and behaviour, spanning a longer period and/or occur at large scale. Related themes can be: shifting precipitation patterns, rising average temperatures, long-term droughts, wildfires.

*GEO** - Geohydrological hazard.
This category applies when the policy element addresses problems rising from mass transfer of water or solid earth, caused by the interaction of atmospheric conditions with land and ocean systems. Related themes can be: floods, landslides, glacial melting, coastal erosion.

*ENV** - Environmental hazard.
This category applies when the policy element addresses problems rising from degradation of ecosystem services and natural systems, posing threat to water, land, air, biodiversity, as well as crucial earth processes. Related themes can be: deforestation, soil degradation, air pollution, desertification.

*BIO** - Biological hazard.
This category applies when the policy element addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna. Related themes can be: vector-borne diseases, invasive species, animal epidemics.

I.5 Sector classification (task 5)

Your task is to identify the sector(s) addressed by a climate change adaptation policy element.

Pick 0, 1, or multiple categories (out of the 5 options) that fit the provided element.
Only select a category when the policy element **explicitly** addresses the sector.
Return an array of the three-letter codes of the categories, without generating any other text.
If no sector category fits the adaptation policy element, return an empty array ([]).

Categories:
*FOO** - Food production sector.
This category applies when the policy element addresses agriculture/crop production, animal and livestock health and productivity, fishery yields, or aquaculture production.

*WAT** - Water resources sector.
This category applies when the policy element addresses availability, accessibility, quality, of water or management of water resources.

*HEA** - Health and wellbeing sector.
This category applies when the policy element addresses air quality, heat, mental health, malnutrition, or infectious diseases.

*BUI** - Built environment sector.
This category applies when the policy element addresses infrastructure, cities, settlements, communities, tourism, or economic sectors.

*NAT** - Natural environment sector.
This category applies when the policy element addresses terrestrial, freshwater, or ocean ecosystems, biodiversity, or phenology.

I.6 Intentionality classification (task 6)

Your task is to classify adaptation policy elements into a **intentionality** class.

Based on the provided policy element, determine whether it intentionally (INT), implicitly (IMP), or indirectly (IND) addresses climate change adaptation.

Consider the following examples:

*Intentional**: flood defense mechanisms, resilience programmes

*Implicit**: climate research, disaster risk reduction

*Indirect**: sustainable development, climate awareness

Return only the three-letter code (INT, IMP, or IND) without returning any other text.

Author Index

Abdi, Reza, 183
Abels, Christoph Maximilian, 282
Adam, Raven, 158
Agaoglu, Orhan, 170
Aggarwal, Vaishali, 298
Ahmed, Shafiuddin Rehan, 183
Aletras, Nikolaos, 51
Alibekov, Nurtilek, 20
Alkaei, Zaher, 282
Athanasiadis, Ioannis N., 315

Bai, Nan, 170
Bao, Alicia, 220
Biesbroek, Robbert, 315
Bonenkamp, Jetske, 315

Chaix, Estelle, 143
Clarke, Christopher, 183
Cohen, Felix, 124
Correia, Ricardo, 199
Creanga, Claudiu, 41

Dasgupta, Tirthankar, 298
de Brito, Mariana Madruga, 1, 99
Deléger, Louise, 143
Dinu, Liviu P, 41

Fernandez, Jeanne, 99
Ferré, Arnaud, 143
Fraser, Alexander, 29

Gironés, Edgar Salas, 170
Gupta, Tanay Kumar, 298

He, Jiamian, 220
Hsu, Angel, 220

Ishisaka, Tatsuya, 116

Jumle, Vedant, 170

Kanchinadam, Teja, 183
Karimi, Sarvnaz, 231
Keuschnigg, Marc, 1
Kogler, Marie Lisa, 158
Kreuter, Frauke, 29
Kumar, Vishal, 246

Laï-king, Mathieu, 143
Le Bodo, Yann, 143
Lioma, Christina, 51
Luan, Tianyu, 183

Madureira, Brielen, 1
Maier, David, 158
Maistro, Maria, 51
Manya, Diego, 220
Metinov, Adilet, 20
Mol, Laurens, 170
Mård, Johanna, 99

Nguyen, Vincent, 231
Niekler, Andreas, 1

Oehler, Simon, 132
Oussalah, Mourad, 246

Parra, Isabel M., 199
Pedrini, Helio, 85

Ray, Georgia, 124
Reichenbach, Lisa, 29, 132
Ruotsalo, Tuukka, 51

Schierholz, Malte, 29, 132
Schulten, Anna Steinberg, 29, 132
Singh, Dhirendra, 231
Siow, Eugene, 183
Soto, Gabriela Morales, 85
Stede, Manfred, 282

Tashbaltaev, Tynchtykbek, 20
Tropeano, Pietro, 51
Turek, Tomas, 231

Uulu, Bektemir Kumarbai, 20
Uvalieva, Zarina, 20

Verma, Ishan, 298

Xie, Yong, 199

Zahra, Shorouq, 99
Zhang, Ji, 220
Zhang, Xiuzhen, 231
Zhang, Yonghong, 199