

Fine-Tuned Encoders versus Zero-Shot Prompting for Detecting Threat and Solution Framing in German Climate News

Raven Adam¹ David Maier¹ Marie Kogler^{1,2}

¹University of Graz, Austria ²Graz 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; Giliardi 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

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 pre-trained 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 $\pm$ 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.

Model	Threat				Solution			
	T	-1	± 1	± 4	T	-1	± 1	± 4
<i>Fine-tuned encoders</i>								
GBERT-large	.760 $\pm$.011	.781 $\pm$.017	.793 $\pm$.013	.805 $\pm$.017	.754 $\pm$.014	.777 $\pm$.013	.792 $\pm$.013	.802 $\pm$.013
GELECTRA-large	.759 $\pm$.013	.775 $\pm$.016	.793 $\pm$.018	.802 $\pm$.019	.750 $\pm$.022	.773 $\pm$.013	.784 $\pm$.014	.794 $\pm$.017
GeistBERT-base	.752 $\pm$.026	.763 $\pm$.016	.775 $\pm$.015	.782 $\pm$.019	.743 $\pm$.013	.765 $\pm$.014	.776 $\pm$.010	.781 $\pm$.012
<i>Zero-shot prompting</i>								
MiMo-V2.6-Flash	–	–	–	.807 $\pm$.014	–	–	–	.798 $\pm$.022
Qwen3.8-27B	–	–	–	.790 $\pm$.010	–	–	–	.777 $\pm$.014
Llama 4 Maverick	–	–	–	.791 $\pm$.016	–	–	–	.770 $\pm$.015
Jev-1.13.0	–	–	–	.701 $\pm$.022	–	–	–	.772 $\pm$.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. Boydston, 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. *Geist-BERT: 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 Stambach, 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.