Retrieval-Augmented LLMs and Encoder Models for Multi-Label Climate Disinformation Narrative Classification

Neda Foroutan[1*], Alexandra Tsiakalou[1*], Vera Schmitt[1,2,3,4]
[1]Technische Universität Berlin
[2]German Research Center for Artificial Intelligence (DFKI)
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[4]Centre for European Research in Trusted AI (CERTAIN)
Berlin, Germany
{neda.foroutan, tsiakalou, vera.schmitt}@tu-berlin.de
[*]These authors contributed equally as first authors.

Abstract

The detection of climate disinformation narratives remains challenging due to label imbalance, hierarchical taxonomies, and the multi-label nature of real-world claims. Developing models that can reliably assign fine-grained narrative categories is therefore essential for scalable analysis of climate disinformation. We present our approach to multi-label climate disinformation narrative classification for ClimateCheck@NSLP 2026 Task 2. The task requires assigning one or more narrative categories, defined by the hierarchical CARDS taxonomy, to climate-related claims. We investigate both encoder-based transformers and decoder-only large language models (LLMs), comparing fine-tuning BERT-based models with prompt-based and retrieval-augmented instruction tuning strategies with Qwen3 model. To address data scarcity and label imbalance, we explore targeted augmentation using external CARDS-based resources as well as semantic similarity filtering. Our experiments show that augmentation improves encoder-based models, with ModernBERT achieving competitive performance at low computational cost. However, the strongest results are obtained using retrieval-augmented instruction tuning with Qwen3, which narrows the candidate narrative space prior to prediction. This approach achieves a Macro-F1 score of 59.72% on the official test set, securing second place on the leaderboard. These findings demonstrate the effectiveness of retrieval-guided LLM adaptation for structured multi-label narrative classification while highlighting the continued relevance of efficient encoder-based models.

Keywords: climate change disinformation, narrative classification, multi-label classification, large language models

1. Introduction

Climate change has become a central topic in online and social media discourse. At the same time, digital platforms enable the rapid dissemination of false or misleading narratives about climate science, which can undermine public trust in scientific institutions and influence policy decisions. Automatically detecting and categorizing climate disinformation narratives has therefore emerged as a critical challenge for natural language processing.

The ClimateCheck@NSLP 2026 shared task (Abu Ahmad et al., 2026) addresses this challenge through two sub-tasks: scientific fact-checking and disinformation narrative detection. In this work, we focus on Task 2, which targets the classification of climate-related claims according to the CARDS taxonomy (Coan et al., 2021). Given a claim, systems must identify whether it expresses one or more disinformation narratives or none making the task a multi-label classification problem in which each claim may be associated with zero, one, or multiple narrative categories.

To tackle this problem, we explore both encoder-based transformer models and decoder-only large language models (LLMs). For encoder models, we fine-tune BERT-based architectures and apply targeted data augmentation to mitigate class imbalance and enhance generalization. For decoder-only models, we investigated instruction tuning with prompt-enhanced, hierarchical instruction tuning, and retrieval-augmented instruction tuning, leveraging the structured nature of the CARDS taxonomy to guide prediction.

Our results demonstrate that retrieval-augmented instruction tuning achieves the highest Macro-F1 scores, outperforming both the official ClimateCheck 2026 baseline and our fine-tuned BERT-based systems. Smaller transformer models remain competitive while requiring substantially lower computational cost and environmental footprint. All scripts and resources are publicly available at: https://github.com/XplaiNLP/ClimateCheck-NSLP-2026.

2. Related Works

Coan et al. (2021) presented one of the earliest efforts to systematically categorize climate change contrarian claims. The authors introduced the CARDS dataset by collecting climate-related text from blogs and conservative think-tank websites and developed a comprehensive contrarian climate

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 225–234
12 May 2026.

taxonomy, structured across three hierarchical levels, known as the CARDS taxonomy. Using this framework, they applied supervised machine learning approaches to classify contrarian climate claims, including logistic regression, linear Support Vector Machines (SVM), a pretrained language model, and RoBERTa-large pretrained transformer language model.

Building on this foundation, Rojas et al. (2024) extended the CARDS framework to the social media domain. They proposed the Augmented CARDS framework, a two-stage hierarchical model designed to detect climate disinformation on Twitter, a domain not covered in the original CARDS study. The authors collected contrarian climate-related tweets over a six-month period in 2022 and additionally incorporated the Climate Change Twitter Dataset released by the University of Waterloo. The combined dataset was annotated according to the CARDS taxonomy, with a refinement that separates category 5.3 from category 5.2. Their modeling approach employed DeBERTa in two sequential stages: first, a binary classifier distinguished contrarian claims from non-narrative content; second, identified contrarian claims were assigned fine-grained CARDS taxonomy labels. Their results demonstrated the effectiveness of hierarchical modeling for large-scale disinformation detection in noisy social media environments.

In addition to taxonomy-driven and supervised approaches, prior work has also explored data augmentation strategies for climate disinformation detection. J et al. (2022) investigated the impact of synthetic data generation techniques to improve the classification of climate change denial narratives. The authors examined augmentation methods such as paraphrasing and controlled text generation to mitigate class imbalance in CARDS dataset, demonstrating that augmented data can improve SVM and RoBERTa-large model robustness in low-resource settings. Their findings highlight the importance of data diversity and label distribution when training disinformation detection models.

Beyond CARDS-based studies, prior work has also examined climate-related narratives in other domains. (Rowlands et al., 2024) studied predicting narratives of Climate Obstruction in Social Media Advertising employs a RoBERTa-large classifier to analyze narrative framing in social media advertisements, using a distinct label schema that differs from the CARDS taxonomy. This illustrates domain-specific variations in climate obstruction narratives.

3. Datasets

The ClimateCheck 2026 dataset consists of 939 unique English claims related to climate change. Each claim is annotated with one or more disinformation narrative labels following the CARDS taxonomy (Coan et al., 2021), which defines a structured hierarchy of climate disinformation narratives. The labels span two levels of the taxonomy: super-claims (top-level categories) and sub-claims (fine-grained categories). As the annotation scheme is multi-label, a single claim may be associated with multiple narratives. The dataset is divided into a training set of 763 claims and a test set of 176 claims. It is available at: https://huggingface.co/datasets/rabuahmad/climatecheck.

Due to the relatively small size of the ClimateCheck training set, we further incorporated two additional resources in our experiments to enhance model performance: the original CARDS dataset (Coan et al., 2021) and the Augmented-CARDS dataset (Rojas et al., 2024), containing 9,067 and 10,661 instances, respectively. These supplementary datasets provide additional annotated climate-related disinformation claims, obtained from contrarian blogs, conservative think tank websites, and Twitter, and annotated under the same CARDS taxonomy. However, in contrast to ClimateCheck, these datasets provide single-label annotations, where each claim is assigned exactly one narrative category.

3.1. Augmented_CC26: Data Augmentation with External Datasets

In order to mitigate the label imbalance present in the ClimateCheck dataset, as well as its limited size, we augmented the dataset using external resources aligned with the CARDS taxonomy. Specifically, we used the Augmented CARDS dataset to enrich underrepresented classes. For every class other than 0_0 (no disinformation narrative present), which was the majority class, we randomly sampled up to 300 examples from Augmented CARDS and added them to the original training data. However, not all rare classes were able to be augmented, as many were also underrepresented or entirely missing from the external dataset, as well. We experimented with different augmentation scales, including sampling maximum 100 examples per class, and fully merging Augmented CARDS with the ClimateCheck dataset. However, for ModernBERT, both strategies resulted in lower Macro F1 validation scores, and were therefore not pursued further. The resulting dataset [1] consists of 6030 claims.

[1] https://huggingface.co/datasets/alextsiak/augmented_cc26

3.2. Exorde: Semantic Similarity-Based Augmentation

Additionally, in some experiments we used a subset of a dataset by Exordia Labs (Exorde Labs, 2024) to further augment the ClimateCheck2026 data according to the CARDS taxonomy. This dataset includes over 260 million social media posts, blog posts, and news articles, captured over a one month period from November to December 2024. Each text includes metadata, such as thematic categorization and a set of English keywords representing the core content of the text.

For our experiments, we first filtered the dataset by thematic categories, selecting posts in English labeled 'Environment', 'Health', 'Science', or 'Politics' that contained the keyword 'climate'. We then computed the semantic similarity of these filtered posts to examples in the Augmented CARDS dataset, using the e5-large-v2 pre-trained embedding model (Wang et al., 2024) and cosine similarity. Only the posts above a similarity threshold of 0.86 were retained. This threshold was selected by inspecting the distribution of cosine similarities across a sample of posts, and manually reviewing high-scoring candidates to identify the point at which retrieved posts were consistently similar in content to known climate disinformation claims. Since many of the retrieved texts were only thematically similar to the Augmented CARDS claims, but did not contain any disinformation, we applied the binary disinformation classifier by Rojas et al. (2024) and kept only posts predicted as disinformation. The resulting dataset contained 1690 posts.

3.3. Label Space Alignment

The ClimateCheck dataset follows a multi-label annotation scheme while the CARDS and Augmented-CARDS datasets provide single-label annotations. To enable joint training across datasets, we unified the label space under a common multi-label representation aligned with the CARDS taxonomy.

For encoder-based fine-tuning, all annotations were converted into multi-hot vectors. Single-label instances from CARDS and Augmented-CARDS were treated as one-hot vectors within this multi-label framework. For decoder-based instruction tuning, we maintained a consistent multi-label output format in the prompt design and included examples demonstrating both single-label and multi-label cases.

This unified representation allows seamless integration of the datasets while preserving taxonomy consistency. However, it implicitly assumes that single-label annotations are exhaustive, which may introduce mild noise if additional implicit narratives are present but unannotated.

4. Methodology

To address the multi-label narrative classification, we explored both encoder-only models (e.g., BERT-based models) and decoder-only large language models, including LLaMA and Qwen3. For all experiments involving Qwen3, we used the `unsloth/Qwen3-8B`. In the rest of this paper, we refer to this model simply as Qwen3 for brevity. Our approaches are described in detail below.

4.1. BERT-based models

We experimented with fine-tuning a variety of BERT-based transformer models for multi-label classification. To handle the label imbalance in the the training data, we additionally explored data augmentation strategies using external datasets.

The main transformer-based model used for our experiments was `ModernBERT-large` (Warner et al., 2024). Moreover, we experimented with `RoBERTa-large` (Liu et al., 2019), `DistilBERT-base-uncased` (Sanh et al., 2020), and `Llama-3.1-8B-Instruct`[2] using LoRA.

Prior to training, narrative labels were converted into multi-hot vectors to support multi-label classification. Input texts were tokenized using the respective model tokenizers with a maximum sequence length of 100 tokens, applying padding and truncation. This relatively short maximum length was chosen because of the concise nature of social media posts and the claims in the ClimateCheck dataset.

We finetuned ModernBERT, RoBERTa and DistilBERT only on the official ClimateCheck 2026 training dataset, utilizing binary cross-entropy loss for all. During training, Macro-F1 score on the validation set was monitored, and early stopping was applied with a patience of two epochs.

We also finetuned ModernBERT on the Augmented_CC26 dataset, as well as Llama-3.1-8B-Instruct using LoRA, to evaluate the effect of targeted augmentation on multi-label narrative classification.

4.2. Prompt Enhancement for Qwen3

We built upon the official ClimateCheck 2026 baseline pipeline [3] and enhanced the original prompt through targeted prompt engineering. Additionally, we augmented the enhanced prompt with three additional in-context examples to better guide the model's predictions. The added examples represent the following scenarios: (i) a claim associated

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://github.com/XplaiNLP/ClimateCheck-2026-Baseline

with multiple narrative IDs (multi-label case), (ii) a claim assigned a single narrative ID, and (iii) a claim labeled as containing no disinformation narrative. This design ensures that the model is exposed to the full range of possible output structures. The final prompt template is shown in Figure 1 (Appendix A.1).

Following the baseline pipeline, we fine-tuned the Qwen3 chat model using supervised instruction tuning with LoRA. Fine-tuning was performed with the final prompt and the data sets described in Section 3.

4.3. Hierarchical Instruction Tuning

We further explored a hierarchical instruction tuning strategy to better leverage the structure of the CARDS taxonomy. In the first stage, we instruction-tuned Qwen3 to predict the top-level narrative category among five coarse-grained classes for the input claim. The prompt followed the same structure as described in the previous approach, with the output format adapted to reflect only the top-level category label. To enable this setup, the original narrative annotations were transformed by extracting their corresponding top-level category. The examples included in the prompt were also updated accordingly, as illustrated in Figure 2 (Appendix A.2).

In the second stage, we instruction-tuned Qwen3 again to predict the fine-grained (sub-level) narrative label. In this setting, the model was provided with (i) the input claim, (ii) the predicted top-level narrative category from the first stage, and (iii) the complete narrative taxonomy list. The prompt examples were modified to reflect this updated input structure.

This two-stage hierarchical approach aims to decompose the prediction task into coarse-to-fine steps, potentially reducing label confusion and improving classification consistency.

4.4. Instruction Tuning with Retrieval Augmentation

To incorporate retrieval-based guidance, we employed the pretrained `cross-encoder/nli-deberta-v3-base` CrossEncoder model to compute semantic similarity scores between each claim and the full list of narratives descriptions defined in the CARDS taxonomy. Based on these scores, we selected the top-10 most similar narratives for each claim, as preliminary experiments showed that this configuration yields better performance than using either top-5 or top-15 retrieved narratives.

The retrieved narratives were then incorporated into a new modified prompt provided to Qwen3. Specifically, the prompt included (i) the input claim, (ii) the top-10 retrieved narrative descriptions, and (iii) the complete narrative taxonomy list. In contrast to the previous prompt-based approaches, no in-context examples were included in this setting, as preliminary experiments showed that adding examples did not improve performance. The prompt template used in this retrieval-augmented setup is illustrated in Figure 3 (Appendix A.3) .

5. Evaluation

5.1. Experimental Setup

All experiments were conducted using two NVIDIA T4 GPUs on the Kaggle platform. Training the Qwen3 model required approximately 2.5 hours per experiment. For training and evaluation of the BERT-based models, we used batch sizes of 16 and finetuned them using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2e-5 and a weight decay of 0.01. Models were trained for up to 15 epochs, with early stopping based on Macro-F1 on the validation set, and with each training run completing in approximately 20 minutes.

Due to hardware constraints, the `unsloth/Qwen3-8B` model was loaded in 4-bit precision (`load_in_4bit=True`), which reduces memory consumption while maintaining performance. The rest of parameters are kept the same as ClimateCheck 2026 baseline pipeline.

5.2. Evaluation Metrics

The task is evaluated using Macro-Precision, Macro-Recall, and Macro-, micro-, and weighted-F1 scores, where Macro-F1 is considered as the score for ranking in ClimateCheck 2026.

Models' performance are assessed using Macro-Precision, Macro-Recall, and F1 scores computed under macro, micro, and weighted averaging schemes. Among these metrics, Macro-F1 serves as the primary evaluation criterion and is used for system ranking in ClimateCheck 2026.

5.3. Results

Table 1 summarizes the performance of our main approaches and data augmentation strategies. For experiments involving CARDS and Augmented-CARDS, non-narrative and duplicated samples were removed, and only narrative-labeled claims were retained for training.

5.3.1. BERT-Based Models

Among the encoder-only models, ModernBERT achieved the strongest overall performance on the ClimateCheck 2026 (CC26) test set, reaching a Macro-F1 score of 41.36% and outperforming both

Model	Macro-Precision	Macro-Recall	Micro-F1	Weighted-F1	Macro-F1
DistilBERT + CC26	0.3194	0.4353	0.7324	0.7212	0.3505
RoBERTa + CC26	0.3512	0.5096	0.7428	0.7524	0.3832
ModernBERT + CC26	0.3860	0.5155	0.7700	0.7653	0.4136
ModernBERT + Aug_CC26 + Exorde	0.4196	0.6283	0.7341	0.7668	0.4583
Qwen3 & prompt + CC26	0.5146	0.5214	0.8333	0.8067	0.4926
Hierarchical + CC26	0.5574	0.5140	0.8189	0.7890	0.4936
ClimateCheck 2026 Baseline	0.5298	0.5736	0.7977	0.7843	0.5136
ModernBERT + Aug_CC26	0.5537	0.5741	0.8235	0.8171	0.5247
Qwen3 & prompt + Aug_CC26	0.5374	0.5918	0.8264	0.8104	0.5431
Qwen3 & Retrieval + CC26	0.5848	0.5699	0.8474	0.8123	0.5450
Llama-8B + Aug_CC26	0.5738	0.5864	0.8441	0.8369	0.5475
Qwen3 & prompt + CC26 + Aug_CARDS	0.5446	0.6135	0.8296	0.8133	0.5547
Qwen3 & prompt + CC26 + Full-CARDS	0.5474	0.6197	0.8351	0.8192	0.5587
Qwen3 & Retrieval + CC26 + Full-CARDS	**0.6249**	**0.6398**	**0.8011**	**0.8061**	**0.5972**

Table 1: Model performance on the ClimateCheck 2026 test set. CC26 is ClimateCheck 2026 dataset. "Full-CARDS" denotes the combination of the CARDS and Augmented-CARDS datasets. Model variants are defined as follows: Qwen3 & prompt: Prompt Enhancement for Qwen3, Hierarchical: Hierarchical Instruction Tuning, Qwen3 & Retrieval: Instruction Tuning with Retrieval Augmentation

RoBERTa and DistilBERT. As a more recently developed transformer model trained on substantially larger and more diverse corpora, ModernBERT appears better suited for complex multi-label classification tasks.

We additionally experimented with a full merge of the CC26 and Augmented-CARDS datasets; however, this approach resulted in decreased validation performance for the BERT-based models. In contrast, targeted augmentation using the Augmented_CC26 dataset improved performance across all encoder models. This suggests that carefully selected additional data can partially mitigate class imbalance in the original training set. Notably, ModernBERT trained with Augmented_CC26 data surpassed the shared-task baseline, demonstrating that compact transformer models can achieve competitive performance without the computational overhead of large language models.

In contrast to the above augmentation strategy, the semantic similarity-based augmentation (Exorde) did not lead to improved performance, decreasing the Macro-F1 score substantially. Manual inspection of a subset of this dataset revealed that, while the binary disinformation classifier effectively filtered non-disinformation posts, the fine-grained narrative labels assigned through semantic similarity to the Augmented CARDS data were often noisy. We observed that 60.4% of claims were classified as 5_1, 5_2, or 5_1. While these labels, along with the 2_0 category and its sub-narratives, seem to be fairly accurate, other narratives (particularly under 1_0 and 4_0) are more frequently misclassified. Table 2 presents examples of such misclassifications. In addition, some misclassifications seem to be due to posts in the Exorde dataset frequently containing multiple core claims. In contrast, the ClimateCheck dataset contains atomic claims; each claim represents a single statement, even if it can be associated with multiple narratives.

This indicates that semantic similarity alone might be insufficient for reliable narrative labeling, though can be more effective when used as contextual guidance, as explored in our LLM-based approaches.

5.3.2. Instruction Tuning with Qwen3

For our instruction-tuned Qwen3 models, prompt enhancement on CC26 data set achieved the Macro-F1 score of 49.26%, which did not surpass the baseline. However, when we incorporated the enhanced prompt approach with augmented data the performance improved up to 6%. The best performance for the Prompt Enhancement approach (55.87% Macro-F1) was achieved when training on the combined CC26, CARDS, and Augmented-CARDS datasets.

The Hierarchical Instruction Tuning approach on CC26 yielded performance comparable to Prompt Enhancement on the same data. This suggests that decomposing the prediction into two stages, first predicting the top-level category and then the sub-level label, did not provide additional benefits compared to directly predicting fine-grained labels in a single step.

The Retrieval Augmented model achieved the strongest performance among all our approaches. When Qwen3 was instruction-tuned with retrieval augmentation using only the CC26 dataset, it reached a Macro-F1 score of 54.50%. Perfor-

Claim	Assigned Label	Suggested Label
'Man-made climate change' IS a thing - it's just not what 'they' tell you it is. They tell you that natural climate change is man-made, which it isn't, and cannot it be controlled. Weather modification however IS man-made, controllable and has been weaponised.	4_5	2_1
As a Canadian and a mom to three I can't believe that our PM @JustinTrudeau would put feeding my family and keeping a roof over our heads a lower than keeping his climate BS going. He's a parent, I doubt his kids go without anything. Absolutely disgusting comments. He taxes us.	4_4	4_2
There is no climate crisis.	3_0	1_0
CLIMATE TERRORISTS...NO CLIMATE CHANGE YES CLIMATE SCAM..	1_8	5_3
THERE IS ZERO EVIDENCE TO LINK CARBON DIOXIDE WITH THESE CLIMATE EVENTS. IT IS SCARE MONGERING AT ITS WORST.	1_7	2_3

Table 2: Examples from error analysis of the Exorde dataset, showing misclassifications by semantic similarity-based labeling and suggested correct labels.

Team	Macro-Precision	Macro-Recall	Macro-F1
ahilbert	0.7071	0.6310	0.6247
XplaiNLP	0.6249	0.6398	0.5972
ClimateSense	0.6700	0.5678	0.5834

Table 3: Comparison of the top three teams on the ClimateCheck 2026 leaderboard, evaluated using Macro-F1 on the official test set.

mance further improved when training was conducted on the combined CC26, CARDS, and Augmented-CARDS datasets, achieving the highest overall Macro-F1 score of 59.72%. These findings suggest that incorporating a retrieval step effectively narrows the candidate narrative space and provides semantically relevant contextual information prior to prediction, thereby substantially enhancing Qwen3's classification performance.

5.3.3. Leaderboard results

On the official ClimateCheck 2026 leaderboard, our best-performing model, instruction tuning of Qwen3 with retrieval augmentation, achieved a Macro-F1 score of 59.72%, securing second place among all participating teams. The *ahilbert* team obtained the highest Macro-F1 score of 62.47%. Table 3 presents a comparison of Macro-Precision, Macro-Recall, and Macro-F1 scores across the teams. While our team, *XplaiNLP*, achieved the second-highest Macro-F1 score, it obtained the highest Macro-Recall.

6. Conclusion

In this work, we investigated multiple approaches for multi-label climate disinformation narrative classification in the context of ClimateCheck 2026 Task 2. We evaluated both fine-tuning encoder-based transformer models and instruction-tuned decoder-only language models, exploring prompt enhancement, retrieval augmentation, and hierarchical prediction strategies. Our findings demonstrate that targeted data augmentation consistently improves model performance and helps mitigate label imbalance in the original dataset. Among all approaches, instruction tuning of Qwen3 with retrieval augmentation achieved the strongest overall results, outperforming both the shared-task baseline and fine-tuned BERT-based models. At the same time, ModernBERT combined with carefully selected augmented data achieved competitive performance while requiring substantially lower computational resources. These results highlight the effectiveness of retrieval-guided instruction tuning of decoder large language models for complex multi-label narrative classification, while also demonstrating that smaller encoder models remain strong and efficient alternatives.

Limitations

Our work is constrained by computational and memory limitations, which restricted the size of the models and the volume of training data that could be used during instruction tuning. In particular, hardware constraints limited our ability to experiment with larger language models, longer training schedules, and more extensive hyperparameter optimiza-

tion.

Acknowledgments

This research was carried out as part of the *VeraXtract* (reference: 16IS24066) and *news-polygraph* (reference: 03RU2U151C) projects, both supported by funding from the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

7. References

Raia Abu Ahmad, Max Upravitelev, Aida Usmanova, Veronika Solopova, and Georg Rehm. 2026. ClimateCheck 2026: Scientific Fact-Checking and Disinformation Narrative Classification of Climate-related Claims. In *Proceedings of the 3rd International Workshop on Natural Scientific Language Processing (NSLP 2026)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Travis G Coan, Constantine Boussalis, John Cook, P. Nanko, and K.M. O'Connor. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320.

Exorde Labs. 2024. Multi-source, multi-language social media dataset. https://www.exordelabs.com/. [Data set].

Piskorski J, Nikolaidis N, Stefanovitch N, Kotseva B, Vianini I, Kharazi S, and Linge J. 2022. Exploring data augmentation for classification of climate change denial: Preliminary study. *CEUR Workshop Proceedings*, 3117:97–109.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuan-Fang Li. 2024. Hierarchical machine learning models can identify stimuli of climate change misinformation on social media. *Communications Earth & Environment*, 5:436.

Harri Rowlands, Gaku Morio, Dylan Tanner, and Christopher Manning. 2024. Predicting narratives of climate obstruction in social media advertising. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5547–5558, Bangkok, Thailand. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

A. Prompt templates for Qwen3 Instruction Tuning

A.1. Prompt Enhancement for Qwen3

Figure 1 depicts the enhanced prompt template used for instruction tuning, as described in Section 4.2. The prompt retains the general classification structure but incorporates refined instructions and in-context examples to better guide the model's predictions. The examples are designed to cover different output scenarios, including multi-label assignments, single-label predictions, and cases with no disinformation narrative.

A.2. Hierarchical Instruction Tuning

Figure 2 illustrates the prompt template used for top-level narrative classification in the first stage of the hierarchical instruction tuning approach described in Section 4.3. The prompt follows the same general structure as the baseline setup, but the output space is restricted to coarse-grained narrative categories. The in-context examples are adapted accordingly to reflect top-level labels.

A.3. Instruction Tuning with Retrieval Augmentation

Figure 3 presents the full prompt template used in the retrieval-augmented instruction tuning described in Section 4.4. The prompt incorporates the input claim, the top-10 retrieved narratives based on semantic similarity, and the complete narrative taxonomy list.

```
f"""You are an expert system for detecting climate change
   disinformation.
You will be given a single claim. Your task is to classify the claim
   by assigning the most appropriate narrative ID(s) from the
   Narratives_list below.

Instructions:
- Select the narrative ID(s) whose description best matches the claim.
- If the claim clearly aligns with more than one narrative, return all
    applicable narrative IDs separated by a semicolon (;).
- If the claim does NOT contain climate change disinformation or does
   not match any listed narrative, return exactly: 0_0.
- Do NOT explain your reasoning.
- Do NOT output anything other than the narrative ID(s).

Narratives_list:
{Narratives_list_str}

Examples:
Example 1
Claim: "According to recent research, we haven't seen any glacier
   melting or sea level rise since 1995."
Narrative IDs: 1_1;1_6

Example 2
Claim: "According to a recent study, there's no link between CO2
   emissions and temperature changes in our lifetime. #ClimateChange #
   Science"
Narrative IDs: 2_3

Example 3
Claim: "Turns out, species that can adapt easily to different
   environments are often the ones that can survive in a wide variety
   of places."
Narrative IDs: 0_0

Now classify the following claim:
Claim: "{claim}"
Narrative IDs: """
```

Figure 1: Enhanced prompt template for Qwen3 instruction tuning.

```
f"""You are an expert system for detecting climate change
   disinformation.

You will be given a single claim. Your task is to classify the claim
   by assigning the most appropriate narrative ID(s) from the
   Narratives_list below.

Instructions:
- Select the narrative ID(s) whose description best matches the claim.
- If the claim clearly aligns with more than one narrative, return all
    applicable narrative IDs separated by a semicolon (;).
- If the claim does NOT contain climate change disinformation or does
   not match any listed narrative, return exactly: 0.
- Do NOT explain your reasoning.
- Do NOT output anything other than the narrative ID(s).

Narratives_list:
{Narratives_list_str}

Examples:

Example 1
Claim: "Over the past century, the Earth's temperature has risen by
   about 0.1$^\circ$C due to CO2."
Narrative IDs: 1;3

Example 2
Claim: "Interesting fact:  Way back when, CO2 levels were high, but
   the sun wasn't as strong. Makes you wonder about the link between
   solar activity and climate, right?"
Narrative IDs: 2

Example 3
Claim: "Turns out, species that can adapt easily to different
   environments are often the ones that can survive in a wide variety
   of places."
Narrative IDs: 0

Now classify the following claim:
Claim: "{claim}"
Narrative IDs:"""
```

Figure 2: Prompt template for top-level narrative classification in hierarchical instruction tuning.

```
f"""You are an expert system for detecting climate change
    disinformation.
You will be given a single claim and a list of 10 narrative labels
    selected from the full narrative inventory because they are most
    semantically similar to the claim.

Your task:
1. Carefully read the claim.
2. Review the provided Similar_Narratives list.
3. Classify the claim by assigning the most appropriate narrative ID(s
    ) from the provided Similar_Narratives list and Full Narrative
    Inventory.

Instructions:
- Select the narrative ID(s) whose description best matches the claim.
- You MUST choose ONLY from the provided Similar_Narratives.
- If the claim clearly aligns with more than one narrative, return all
     applicable narrative IDs separated by a semicolon (;).
- If the claim does NOT contain climate change disinformation or does
    not match any listed narrative, return exactly: 0_0.
- Do NOT explain your reasoning.
- Do NOT output anything other than the narrative ID(s).

Full Narrative Inventory:
{Narratives_list_str}

Now classify the following claim:
Claim: "{claim}"
Similar_Narratives: {Similar_Narratives_str}
Narrative IDs: """
```

Figure 3: Prompt template for Qwen3 instruction tuning given claim and its top-10 similarity with narratives Taxonomy.