XplaiNLP @ ClimateCheck 2026 Task 2: Comparing Hierarchical Approaches for Fine-Grained Climate Disinformation Narrative Classification

Arthur Hilbert[1,3], **Jing Yang**[1,2,4], **Vera Schmitt**[1,2,3,4]
[1]Technische Universität Berlin
[2]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[3]German Research Center for Artificial Intelligence (DFKI)
[4]Centre for European Research in Trusted AI (CERTAIN)
{arthur.hilbert, jing.yang, vera.schmitt}@tu-berlin.de

Abstract

We present our submission to Task 2 of the ClimateCheck 2026 shared task on Disinformation Narrative Classification which requires assigning climate-contrarian claims to fine-grained disinformation narratives. Using `Qwen3-8B` as a fixed backbone, we systematically compare **data augmentation**, **prompt engineering** and **reinforcement learning**. Our experiments show that structured reasoning, particularly a chain-of-thought (CoT) prompting strategy aligned with the taxonomy's hierarchical structure, substantially improves Macro-F1 over both zero-shot baselines and augmentation-based fine-tuning. Our best configuration achieves ∼**0.625 Macro-F1**, ranking first in Task 2. Our findings demonstrate that carefully designed hierarchical prompting can rival more complex training interventions in low-resource, highly imbalanced narrative classification settings.

Keywords: disinformation, climate change, narrative classification

1. Introduction

As digital platforms play an increasingly central role in shaping perceptions of scientific issues, there is a pressing need for NLP systems that can identify and contextualize questionable claims and connect them reliably to evidence. We present our contribution to Task 2 of Abu Ahmad et al. (2026)'s ClimateCheck 2026 shared task[1] on **Disinformation Narrative Classification**. The shared task addressed the growing challenge of climate-related discourse on social media, where increased public engagement is accompanied by the spread of misleading narratives that emerge, adapt, and recombine over time within online information ecosystems (Abu Ahmad et al., 2025b,a).

The 2026 edition of ClimateCheck combines the previous iteration's tasks into Task 1: Abstract Retrieval and Claim Verification and adds a new Task 2: Disinformation Narrative Classification. However, we only participated only in Task 2 of the shared task, which was centered around assigning claims to fine-grained disinformation narratives derived from the CARDS taxonomy of contrarian climate claims (Coan et al., 2021; Rojas et al., 2024). This problem setting is technically challenging due to overlapping narrative themes, imbalanced label distributions, and a scarcity of data (Abu Ahmad et al., 2025b). In the official evaluation, systems were ranked by Macro-F1 rewarding the accurate classification of rare labels. Our best submission achieved the first place in Task 2 with a score of ∼**0.625 Macro-F1**.

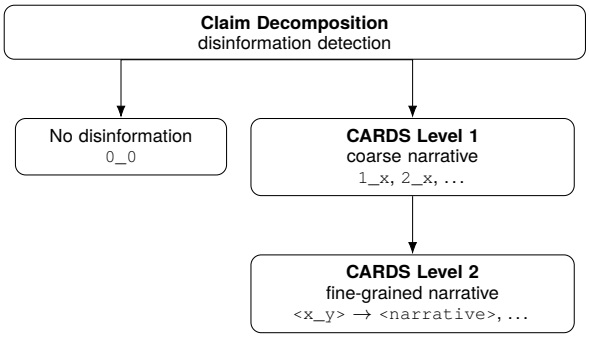

Figure 1: Our three-step hierarchical narrative classification approach.

Rather than varying model families, we focused on isolating the effects of different approaches by using a `Qwen3-8B` model (Yang et al., 2025) as the starting point across experiments. This choice allowed for direct comparison to the task baseline, which performed supervised finetuning on `Qwen3-8B` using Low-Rank Adaptation (LoRA)(Hu et al., 2021). The baseline was provided by team EFC, ClimateCheck 2025's 3rd place submission to the Claim Verification task (Upravitelev et al., 2025). Following the optional shared-task guidance, we tracked the inference emissions of our experiments using CodeCarbon (Courty et al., 2024).

We explore three main directions: **data augmentation**, **prompt engineering**, and **reinforcement learning** using Group-Relative Policy Optimization (GRPO) (Shao et al., 2024). For prompt de-

[1]https://github.com/Dagobert42/climatecheck2026-task2-hierarchical-approaches

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 289–296
12 May 2026.

sign, we investigate *hierarchical prompting* (Singh et al., 2025) and an off-shoot *chain-of-thought (CoT) prompting* (Wei et al., 2023), leveraging the layered structure of the CARDS taxonomy. Our best-performing configuration was achieved with a **CoT prompting** strategy on the vanilla `Qwen3-8B`.

Related Work. ClimateCheck 2026 Task 1 bears much resemblance to the AVeriTeC shared task, where participants must handle evidence retrieval and veracity prediction (Schlichtkrull et al., 2023, 2024). Thus, we searched for comparable shared tasks for Task 2. Our solution leans on the results from SemEval Task 10 Subtask 2, which featured a narrative classification setting within the PolyNarrative dataset (Nikolaidis et al., 2025), but focused on online news articles as opposed to claims (Piskorski et al., 2025). Teams explored approaches from simple vectorization techniques (Blombach et al., 2025) and cross-lingual knowledge transfer (Eleftheriou et al., 2025) to zero-shot agentic frameworks (Eljadiri and Nurbakova, 2025), with prompt- and fine-tuning-based approaches being omnipresent (Piskorski et al., 2025). The leading submission by team GateNLP used data augmentation and introduced a Hierarchical Three-Step Prompting (H3Prompt) approach (Singh et al., 2025). This aligns with earlier work on climate disinformation which also successfully explored data augmentation and hierarchical models (Piskorski et al., 2022; Rowlands et al., 2024; Rojas et al., 2024). Previously overlooked seems to have been the application of reinforcement learning, which has successfully been applied to mathematics (Yang et al., 2024; Wang et al., 2024) and other reasoning-intensive tasks (Suma and Dauncey, 2025).

2. Task Description.

Given a claim c and a predefined taxonomy of climate disinformation narratives $\mathcal{Y} = \{y_1, \ldots, y_M\}$, the task is to predict the subset of labels $\hat{\mathcal{Y}}_c \subseteq \mathcal{Y}$ that apply to c. This is a multi-class, multi-label classification problem, where each claim may be associated with multiple narratives. The goal is to approximate the gold label set $\mathcal{Y}_c^*$ for each claim.

This is commonly modeled as a multi-stage hierarchical classification (Singh et al., 2025; Eleftheriou et al., 2025), where each step is meant to narrow down the prediction to an increasingly reduced set of labels.

Dataset. The dataset at hand is highly imbalanced and comprised a label set of 33 narratives. The training split contained 763 distinct claims, while the test split consisting of 172 claims was used exclusively for submissions. Each datapoint consists of a claim with ID and its narrative label(s).

Narrative Label		Count
0_0	No disinformation narrative	556
5_1	Climate science is uncertain	44
2_1	It's natural cycles/variation	38
...		
1_8	Doesn't impact health	1
2_2	It's non-greenhouse gas forcings	1
1_5	Oceans are not warming	1

Table 1: Most and least frequent narratives in the train dataset.

3. Methods

We compare approaches pertaining to prompting and fine-tuning large language models (LLM). Models were sourced through the `Transformers` library (Wolf et al., 2020) and fine-tuned using `Unsloth` (Daniel Han and team, 2023).

Data Augmentation. Following Singh et al. (2025), we perform targeted data augmentation using (`qwen3-30b-a3b-instruct-2507`) to mitigate severe class imbalance and enrich rare narratives. We first apply a skewed inverse-frequency reweighting to oversample minority narratives from the training set. Given the inverse-frequency w_i for each label, we obtain its sampling probability p_i via normalization and apply a smoothing factor α to maintain some of the original distribution:

$$p_i = \frac{w_i^\alpha}{\sum_{j=1}^{N} w_j^\alpha} \qquad \text{with } \alpha = 0.5$$

With this we sample a list of 1,000 augmentation candidates with repetition and subsequently employ a two-step prompting strategy to improve augmentation quality. Initially, a model is asked to generate a concise explanation of why a candidate claim fits its assigned narrative label(s). Secondly, given narrative labels and the generated explanation, the model produces 10 new synthetic claims for the same narrative label(s). We sampled 10,000 synthetic claims under this procedure, which were subsequently used in a supervised pre-training objective.

We perform parameter-efficient supervised fine-tuning of using LoRA adapters with rank $r = 16$, $\alpha = 16$. The original training data is split further into an 80% train and a 20% validation set. The model is pre-trained on the augmentation data with the validation loss as an early stopping criterion at a patience of 3. Then the model is further fine-tuned using the same criterion on the 80% train split. Finally, we perform a short 2-epoch tuning on the validation split to get maximum mileage out of the limited train data.

Experiment	Macro-F1	Macro-P	Macro-R	Micro-F1	Weighted-F1
Baseline	51.36	52.99	57.37	79.78	78.44
Zero-shot	42.85	45.51	50.98	74.51	73.54
Zero-shot + CoT	38.24	39.59	46.50	68.38	69.33
Zero-shot + H	43.13	39.37	58.18	58.01	62.50
Zero-shot + Reasoning	56.19	61.46	58.23	84.18	80.75
Zero-shot + CoT + Reasoning	**62.48**	**70.71**	**63.11**	**84.42**	82.06
A + FT	54.18	55.18	57.96	84.33	**82.20**
H + A + FT	54.22	57.97	56.84	83.57	81.86
RL	57.20	64.91	59.60	84.33	81.68
CoT + RL	60.47	70.21	58.80	83.05	81.56

Table 2: Performance metrics across experiments using `Qwen3-8B` as base model. All values in %. Best results (highest) per metric are highlighted in **bold**. *Baseline* results were provided by organizers via fine-tuning `Qwen3-8B` on task training data. Zero-shot = no additional training was applied, A = additional pre-training using augmented data, FT = fine-tuning on original training data, H = hierarchical prompting, CoT = chain-of-thought prompting, RL = additional fine-tuning with reinforcement learning using GRPO.

Prompt Engineering. We compare three prompting strategies:

(1) Simple A direct instruction-based prompt asks the model to classify a claim according to the taxonomy without explicitly structuring the reasoning process.

(2) Hierarchical The classification is decomposed into a two-turn conversation, where first, the model identifies the high-level narrative group(s). Second, conditioned on this prior selection, it predict the fine-grained sub-label(s). This explicitly operationalizes hierarchical classification.

(3) CoT To arrive at its prediction the model is instructed to follow an implicit hierarchical process with three steps: (i) extracting the core assertion(s) and separating non-disinformation claims, (ii) selecting the appropriate high-level narrative group (1_x to 5_x), and (iii) choosing the most specific sub-label. This encourages hierarchical reasoning while maintaining a single forward pass.

We provide an overview of the exact prompt templates in the Appendix A.

GRPO Fine-Tuning. We experiment with further reasoning optimizations using GRPO as implemented in the `TRL` library (von Werra et al., 2020). The base model was prepared using LoRA (Hu et al., 2021) with rank $r = 32$, $\alpha = 32$. We train for 250 steps with a learning rate of 5×10^{-6} and 8 sampled generations per prompt.

Claims for GRPO fine-tuning are sampled using the same skewed inverse-frequency reweighting as for data augmentation (with $\alpha = 0.5$). GRPO uses a weighted and signed reward function, which simply assigns the inverse class-weight for correct predictions and an equal negative weight for incorrect predictions. In addition, a brevity reward penalizes overly long generations beyond 250 words to prevent excessive reasoning.

4. Results

We report classification metrics and environmental impact for all evaluated approaches on Task 2. Table 2 summarizes predictive performance.

Experiment	Time	Emissions	Energy
	s	$kgCO_2eq$	kWh
ZS	187	0.0053	0.0138
ZS + CoT	**119**	0.0055	0.0145
ZS + H	700	0.0352	0.0924
ZS + Reasoning	4838	0.1489	0.3910
ZS + CoT + Reasoning	2438	0.1296	0.3403
A + FT	206	0.0090	0.0237
H + A + FT	160	**0.0045**	**0.0118**
RL	2458	0.1261	0.3310
CoT + RL	2541	0.1229	0.3226

Table 3: Emission statistics for the evaluated experiments at inference time. Best (lowest) values are in **bold**. ZS = Zero-Shot setting.

Table 3 provides the corresponding inference-time environmental statistics, including duration (in seconds), estimated emissions ($kgCO_2eq$), and energy consumption (kWh), measured using CodeCarbon under identical hardware conditions.

Figure 2 visualizes the trade-off between predictive performance (Macro-F1) and computational cost (inference duration), mostly induced by use of reasoning capabilities.

5. Discussion

Performance Trends. Across all approaches, structured reasoning substantially improves zero-shot performance. In terms of Macro-F1, naive zero-shot approaches lag considerably behind fine-tuning and structured reasoning methods. Intro-

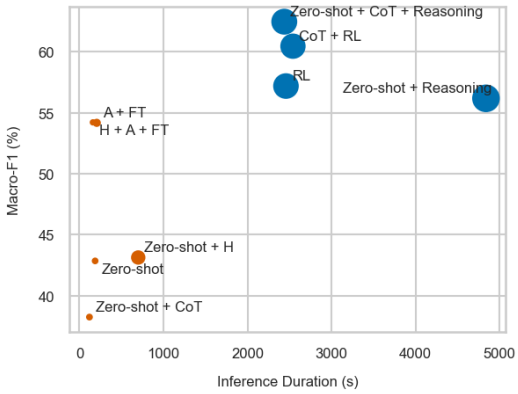

Figure 2: Macro-F1 (%) versus inference duration (s). Blue denotes reasoning-based approaches, orange non-reasoning approaches. Point size proportionally indicates emissions in $\mathrm{kgCO_2eq}$.

ducing explicit reasoning steps yields consistent gains, with the highest overall Macro-F1 achieved for CoT prompting with reasoning. This suggests that guided intermediate reasoning helps the model better differentiate fine-grained narrative categories, particularly in a highly imbalanced multi-label setting.

Hierarchical prompting as an intermediate step between simple zero-shot and reasoning methods does not yield significant improvements. Nor does its augmented and fine-tuned variant deviate in any significant way from non-hierarchical augmentation and fine-tuning. Augmented fine-tuning improves over the fine-tuning baseline and yields strong Weighted-F1, reflecting improved alignment with the dominant class. However, Macro-F1 gains remain moderate compared to reasoning-based zero-shot methods. This indicates that synthetic data improves robustness but fails to resolve minority-class discrimination.

Reinforcement learning via GRPO improves reasoning performance for the naive prompt setting but surprisingly falls short when using the CoT-prompt variant. These results remain non-conclusive on how reward shaping with class-sensitive weighting affects minority class predictions.

Efficiency vs Performance Trade-off. The environmental analysis reveals a clear and expected trade-off between predictive quality and computational cost. Pure zero-shot and lightweight prompting strategies are highly efficient but underperform in Macro-F1. In contrast, reasoning-heavy approaches dramatically increase inference time and emissions. Notably, augmentation with subsequent fine-tuning achieves competitive performance with comparatively low emissions. Figure 2 illustrates this Pareto-like frontier.

Error Patterns. As the test set is not made public we cannot provide any confusion matrices. However, minority labels with extremely low training frequency remain challenging across all methods, indicating that imbalance persists even after augmentation and reward weighting.

Limitations. First, Macro-F1 remains sensitive to extremely rare labels, some of which contain only one or two instances. Second, reasoning-based methods incur substantial computational overhead, which inhibits performance comparison across multiple runs to show how stable (or not) our experimental results are. Third, synthetic augmentation relies on model-generated claims which introduces distributional artifacts that do not fully reflect real-world disinformation patterns. Finally, reinforcement learning rewards are label-exact and do not explicitly account for partial or hierarchical correctness, potentially underestimating semantically close predictions.

Overall, our results highlight the importance of structured reasoning and class-aware optimization for fine-grained narrative classification, while underscoring the need to balance predictive gains with environmental and computational cost.

6. Conclusion

We presented our contribution to Task 2 of ClimateCheck 2026 on Disinformation Narrative Classification. Using `Qwen3-8B` as a fixed backbone, we systematically compared data augmentation, prompt engineering, and reinforcement learning. Our results demonstrate that structured reasoning, particularly using a CoT-based prompting strategy, improves Macro-F1, ultimately achieving first place in the shared task with $\sim$0.625 Macro-F1.

Beyond predictive performance, we evaluated the environmental cost of each method. Our findings contextualize a stark trade-off between reasoning-based gains and computational overhead, highlighting the importance of reporting emissions alongside accuracy in shared tasks. While augmentation and GRPO can provide additional improvements, they do not consistently outperform carefully constructed prompting strategies in this low-resource setting.

Future work should explore the stability of predictive performance across multiple inference runs, hybrid reward formulations to account for hierarchical correctness, and techniques for reducing the computational footprint of reasoning-heavy approaches.

Acknowledgements

The work on this paper is performed in the scope of VeraXtract (16IS24066) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

Bibliographical 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).

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025a. The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 42–56, Vienna, Austria. Association for Computational Linguistics.

Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025b. The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 263–275, Vienna, Austria. Association for Computational Linguistics.

Andreas Blombach, Bao Minh Doan Dang, Stephanie Evert, Tamara Fuchs, Philipp Heinrich, Olena Kalashnikova, and Naveed Unjum. 2025. Narrlangen at SemEval-2025 task 10: Comparing (mostly) simple multilingual approaches to narrative classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2240–2248, Vienna, Austria. Association for Computational Linguistics.

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

Benoit Courty, Victor Schmidt, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, SabAmine, inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Amine Saboni, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, alencon, Michał Stęchły, Christian Bauer, Lucas-Otavio, JPW, and MinervaBooks. 2024. mlco2/codecarbon: v2.4.1.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Konstantinos Eleftheriou, Panos Louridas, and John Pavlopoulos. 2025. KostasThesis2025 at SemEval-2025 task 10 subtask 2: A continual learning approach to propaganda analysis in online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 899–908, Vienna, Austria. Association for Computational Linguistics.

Mohamed Nour Eljadiri and Diana Nurbakova. 2025. Team INSALyon2 at SemEval-2025 task 10: A zero-shot agentic approach to text classification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 965–980, Vienna, Austria. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Nikolaos Nikolaidis, Nicolas Stefanovitch, Purificação Silvano, Dimitar Iliyanov Dimitrov, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ion Androutsopoulos, Preslav Nakov, Giovanni Da San Martino, and Jakub Piskorski. 2025. PolyNarrative: A multilingual, multilabel, multidomain dataset for narrative extraction from news articles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31323–31345, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alipio Mario Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Guimaraes, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval 2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2610–2643, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, and Jens P. Linge. 2022. Exploring data

augmentation for classification of climate change denial: Preliminary study. In *Proceedings of Text2Story - Fifth Workshop on Narrative Extraction From Texts held in conjunction with the 44th European Conference on Information Retrieval (ECIR 2022), Stavanger, Norway, April 10, 2022*, volume 3117 of *CEUR Workshop Proceedings*, pages 97–109. CEUR-WS.org.

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(1):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.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2025. Gatenlp at semeval-2025 task 10: Hierarchical three-step prompting for multilingual narrative classification.

Adam Suma and Samuel Dauncey. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Max Upravitelev, Nicolau Duran-Silva, Christian Woerle, Giuseppe Guarino, Salar Mohtaj, Jing Yang, Veronika Solopova, and Vera Schmitt. 2025. Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 281–287, Vienna, Austria. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement.

A. Appendix

We provide the exact prompt templates used in our experiments for data augmentation and prompt-based classification.

Explanation Prompt This prompt generates a concise explanation of why a claim fits specific disinformation narratives. It is used as an intermediate step in the data augmentation pipeline to improve the quality of synthetic samples.

```
Your task is to concisely explain why
    the following claim fits specific
disinformation narrative(s) on climate
    change.

Claim: {claim}
Narrative IDs: {narrative_codes}
Narrative descriptions:
- {narrative_descs}

Provide a short reasoning of why the
    claim caters to the specific
narrative(s). Do not repeat the claim
    and output nothing else.
Start directly with the explanation and
    limit your response to one
    paragraph.
```

Augmentation Prompt This prompt generates synthetic training examples conditioned on a claim, its narrative labels, and an explanation. It is used to create additional labeled data for fine-tuning.

```
Your task is to augment training data
    for a classifier to combat climate
change disinformation. Claims in the
    dataset are categorized by the
following taxonomy:

{taxonomy}

Example claim: {claim}
Narrative ID: {narrative_codes}
Narrative descriptions:
    {narrative_descs}
Explanation: {explanation}

Generate 10 new claims that cater ONLY
    to the described disinformation
narrative(s) on climate change.
Each new claim must be distinct and
    consistent with the narrative(s).
Do not include any extra keys,
    commentary, emojis or markdown.

Output ONLY a JSON list in the
    following format:
[
  {"claim": "...", "narrative": "..."},
  ...
]
```

Simple Classification Prompt This prompt performs direct zero-shot classification without explicitly guiding the reasoning process.

```
You are an expert in detecting climate
    change related disinformation.

Your task is to classify the claim
    using the provided taxonomy.

Taxonomy:
{taxonomy}

Rules:
- Output ONLY a valid JSON list of
    strings.
- Each item MUST be exactly "key ->
    value".
- If no disinformation is found, output:
  ["0_0 -> No disinformation narrative"]

Claim: "{claim}"
Output:
```

Basic Reasoning Prompt This prompt performs direct classification without explicitly guiding intermediate reasoning steps. It serves as a lightweight baseline for zero-shot classification.

```
You are an expert in detecting climate
    change related disinformation.
You get a claim and your task is to
    classify the claim using the
    taxonomy
codes provided.

Rules:
- Output ONLY a valid JSON list of
    strings.
- Each item MUST be exactly "key ->
    value" where key is a taxonomy code
  and value is its narrative name.
- If no disinformation is found, output:
  ["0_0 -> No disinformation narrative"]
- Do not output anything else.

Taxonomy:
{taxonomy_str}

Claim: "{claim}"
Output:
```

CoT-style Reasoning Prompt This prompt introduces structured reasoning by guiding the model through a three-step process: extracting claims, identifying high-level narratives, and selecting fine-grained labels.

```
You are an expert in detecting climate
    change related disinformation.

Your task is to classify the claim
    using the provided taxonomy.
```

```
Taxonomy:
{taxonomy}

Follow this 3-step reasoning process
    internally:

Step 1 - Extract the core assertion(s)
Step 2 - Identify the high-level
    narrative group
Step 3 - Choose the most specific
    sub-label

Output ONLY a valid JSON list of
    strings.

Claim: "{claim}"
Output:
```

Hierarchical Prompt (Level 1) This prompt represents the first stage of hierarchical classification, where the model predicts coarse narrative groups.

```
You are an expert in detecting climate
    change related disinformation.

Your task is to classify the claim
    using the provided taxonomy.

Taxonomy:
{taxonomy}

Step 1 - Extract the core assertion(s)
Step 2 - Identify the high-level
    narrative group(s)

Claim: "{claim}"
Output:
```

Hierarchical Prompt (Level 2) This prompt represents the second stage of hierarchical classification, where the model refines predictions into fine-grained narratives.

```
Now complete step 3 using your answers
    from step 2:

Step 3 - Choose the most specific
    sub-label

Output ONLY a valid JSON list of
    strings.

Claim: "{claim}"
Output:
```