Stochastic Gradient Descenders at SemEval-2026 Task 9: Few-Shot LLM Prompting for Polarization Type Classification

Huynh Nguyen Phu and Dang Van Thin
University of Information Technology, Ho Chi Minh, Vietnam
Vietnam National University Ho Chi Minh City, Vietnam
24521352@gm.uit.edu.vn and thindv@uit.edu.vn

Abstract

This paper presents our system for SemEval-2026 Task 9 (POLAR), Subtask 2, which focuses on classifying polarization types in social media text. We investigate three paradigms: (i) fine-tuning mDeBERTa-v3 with domain-adaptive pre-training, (ii) parameter-efficient adaptation of Qwen2.5-32B using LoRA, and (iii) few-shot prompting with Llama-3.3-70B-Instruct. Experimental results show that few-shot prompting, despite requiring no task-specific training, outperforms both fine-tuning and parameter-efficient approaches. Notably, it achieves non-zero F1 scores across all polarization categories, which is critical under macro-averaged evaluation. Our system ranks 2nd out of 29 English submissions on the official leaderboard, achieving an F1 Macro of 0.5157. These findings highlight the effectiveness of large instruction-tuned models in low-resource, label-imbalanced classification settings.

1 Introduction

Online polarization—the tendency of public discourse to fragment into opposing and often antagonistic groups—has emerged as a central challenge in social media analysis. The POLAR shared task at SemEval-2026 (Naseem et al., 2026a,b) addresses this issue across 22 languages, requiring systems not only to detect polarization but also to identify its type. In this work, we focus on Subtask 2: Polarization Type Classification, formulated as a multi-label problem where each instance may be assigned zero or more labels from five categories: political/ideological, racial/ethnic, religious, gender/sexual, and other. Our submission targets the English track.

A key challenge of this task lies in the highly imbalanced label distribution, where several categories are extremely underrepresented and a large portion of instances carry no polarization labels. Under macro-averaged F1 evaluation, failure to identify rare categories leads to disproportionately large performance degradation, making robust modeling of low-frequency labels essential.

To investigate how different modeling paradigms address this challenge, we evaluate three approaches spanning a spectrum of supervision levels: (i) fine-tuning mDeBERTa-v3-base (He et al., 2023) with domain-adaptive pre-training (DAPT) (Gururangan et al., 2020) and Asymmetric Loss (Ben-Baruch et al., 2021); (ii) parameter-efficient fine-tuning of Qwen2.5-32B-Instruct (Yang et al., 2024) using LoRA (Hu et al., 2022); and (iii) few-shot prompting with Llama-3.3-70B-Instruct (Grattafiori et al., 2024), which requires no parameter updates. Our findings reveal a consistent and somewhat counterintuitive trend: few-shot prompting with a large instruction-tuned model outperforms both fully fine-tuned and parameter-efficient approaches. In particular, fine-tuned models fail to produce true positives for the rarest category (gender/sexual), yielding zero F1, whereas the prompting-based approach successfully identifies all label types, including low-frequency ones. This capability proves decisive under macro-F1 evaluation and highlights the advantages of large language models in label-imbalanced, low-resource settings.

2 Related Work

Polarization detection shares common ground with several established NLP tasks—hate speech detection, stance classification, and sentiment analysis (Barbieri et al., 2020)—but differs in an important way: the task is *multi-label and fine-grained*. A single text can be simultaneously political and racial in nature, and several of the target categories appear only rarely in the training data.

Large language models can rival specialized fine-tuned models in few-shot settings (Brown et al., 2020), while multilingual encoders such

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 938–942

as mDeBERTa-v3 (He et al., 2023) remain strong when paired with domain-adaptive pre-training (Gururangan et al., 2020) and imbalance-aware losses (Lin et al., 2017; Ben-Baruch et al., 2021). Parameter-efficient methods such as LoRA (Hu et al., 2022) offer a third path, adapting billion-parameter models with modest compute. Our contribution is an empirical comparison of these paradigms under Macro-F1, where covering rare labels can matter more than optimizing frequent ones.

3 System Overview

We explore three approaches, each occupying a different point on the spectrum between task-specific adaptation and reliance on what a model already knows.

3.1 Approach 1: mDeBERTa-v3 + DAPT

Our first approach is a conventional fine-tuning pipeline. Starting from mDeBERTa-v3-base, we run domain-adaptive pre-training—masked language modeling on the full multilingual POLAR corpus across all 22 languages—and then fine-tune for multi-label classification with Asymmetric Loss ($\gamma_{\text{neg}} = 2.0$, $\gamma_{\text{pos}} = 0.5$, clip$= 0.05$). During preprocessing, we apply NFKC normalization, convert emoji into textual descriptions, and replace URLs and usernames with placeholder tokens. Classification thresholds are tuned per label using a blend of global and per-language optimal values. Training runs for 10 epochs with a learning rate of 2×10^{-5} and a maximum sequence length of 128 tokens on a single NVIDIA L4 GPU via Google Colab Pro.

3.2 Approach 2: Qwen2.5-32B + LoRA

Our second approach asks whether a large generative model, lightly adapted to the task, can do better with rare labels. We apply LoRA ($r = 64$, $\alpha = 128$) to all attention and feed-forward projection matrices of Qwen2.5-32B-Instruct and train with SFTTrainer through the Unsloth framework[1] for 3 epochs. Two prompt variants are tested: a *minimal* version providing only the label names, and a *detailed* version that includes full definitions of each polarization category. At inference, the model produces a structured JSON output via greedy decoding. All training and inference for this approach runs on a single NVIDIA H100 80 GB GPU provisioned through FPT AI Cloud.[2]

3.3 Approach 3: Few-Shot Prompting (Final)

Our submitted system takes the most minimal approach of the three: **no training at all**. We simply prompt Llama-3.3-70B-Instruct in a few-shot setting, relying entirely on the model's existing knowledge and in-context learning ability.

Prompt design. Each prompt has four components: (1) a system message that frames the model as a conservative polarization classifier, telling it to default to 0 when in doubt; (2) detailed natural-language definitions for all five categories; (3) 15 few-shot examples from the training set; and (4) the target text. The model returns a JSON object mapping each label to 0 or 1.

Example selection. To ensure balanced label coverage, we sample 5 positive examples (one per polarization type) and 10 non-polarized examples for each query, shuffling their order to mitigate position bias.

Inference. We query the model through FPT Cloud Marketplace's OpenAI-compatible API[3] with temperature $=$ 0.1—chosen as a near-deterministic setting for stable JSON classification—and 10 concurrent workers. Any malformed response is treated as an all-zero prediction.

4 Experimental Setup

The POLAR Subtask 2 corpus (Naseem et al., 2026b) spans 22 languages with 73,681 training, 3,687 development, and 33,288 test instances in total. For the English track, these numbers are 3,222 / 160 / 1,452 respectively. All results are reported in terms of **F1 Macro**, the official ranking metric, alongside per-label F1 scores to illuminate where individual approaches succeed or fail. The official development set was used to choose among the three system families, not to run a broad sweep over few-shot seeds, shot counts, temperatures, or prompt variants; the submitted few-shot configuration fixes seed 42, 15 demonstrations, and temperature 0.1 for reproducibility.

5 Results

Table 1 compares development-set performance across all six configurations, and the picture it

[1] https://github.com/unslothai/unsloth

[2] https://ai.fptcloud.com

[3] https://marketplace.fptcloud.com

Table 1: English dev F1 per label. "—" = 0.000. †Final submission. Only the Llama system achieves non-zero F1 on all five labels.

System	Macro	Pol	R/E	Rel	G/S	Oth
Fine-tuned encoder (mDeBERTa-v3 + DAPT)						
Run 1	.298	.685	.519	.286	—	—
Run 2	.315	.711	.467	.400	—	—
Run 3	.354	**.732**	.438	**.600**	—	—
LoRA fine-tuned LLM (Qwen2.5-32B)						
Detailed prompt	.372	.710	.348	**.600**	—	.200
Minimal prompt	.318	.554	**.533**	.500	—	—
Few-shot prompting (Llama-3.3-70B)						
15-shot†	**.378**	.598	.465	.353	**.400**	**.074**

paints is revealing. The fine-tuned models—both encoder-based and LoRA-adapted—perform well on the majority categories: mDeBERTa Run 3, for instance, reaches .732 on political and .600 on religious, the highest single-label scores in the table. Yet not one of the five fine-tuned configurations manages to score above 0.0000 on gender/sexual. With only 72 positive gender/sexual examples in the 3,222-instance training set (2.2%) and just 3 in the 160-instance dev set, these models fail to recover any true positives for this label. The Qwen-LoRA variants do emit a few gender/sexual positives (2–4 on dev), but all are false positives, suggesting rare-label suppression or miscalibration under the tested training and decoding setup rather than a single encoder-specific artifact.

The Llama few-shot system behaves quite differently. Its scores on individual labels are often lower than those of the fine-tuned models—it sacrifices some precision on the frequent categories in exchange for the ability to recognize the rare ones. What matters, though, is that it is the *only* system to produce non-zero F1 on all five labels, which earns it the highest overall F1 Macro (0.3780). When the evaluation metric averages across labels, even a single zero wipes out the benefit of strong performance elsewhere; covering every category, however imperfectly, matters more than excelling on just a few. This observation is what led us to select the few-shot system as our final submission.

The official leaderboard (as shown in Table 2) supports this observation: our system ranks **2nd out of 29 English submissions**, achieving a macro-F1 score of 0.5157. This result is 0.0165 points below the top-ranked system and substantially outperforms the organizer baseline (0.3333).

A closer examination of per-label performance

Table 2: Official English leaderboard (top 5 of 29 submissions).

Rank	Team	F1 Macro
1	UTokyo Tsuruoka Lab	0.5322
2	**Stochastic Gradient Descenders (Ours)**	**0.5157**
3	NYCU-NLP	0.5135
4	PolaFusion	0.5035
5	Sagarmatha	0.5035
–	POLAR Baseline	0.3333

Table 3: Per-label test results. All labels except Gender/Sexual show higher recall than precision.

Label	P	R	F1
Political	0.688	0.832	0.753
Racial/Ethnic	0.437	0.732	0.547
Religious	0.414	0.706	0.522
Gender/Sexual	0.548	0.515	0.531
Other	0.183	0.293	0.225

(Table 3) reveals the underlying trends. The political/ideological category, which is the most prevalent, attains the highest F1 score (0.753), driven by strong recall (0.832), indicating a tendency to favor recall over precision for this label. Similar recall-oriented behavior is observed for the racial/ethnic and religious categories.

Notably, the gender/sexual category—where all fine-tuned models failed to produce true positives on the development set—achieves an F1 score of 0.531 on the test set, providing strong empirical support for our model selection strategy. In contrast, performance on the "Other" category remains limited (F1 = 0.225), likely due to its broad and ambiguous definition, which makes it difficult to model precisely.

6 Error Analysis

To better understand the system's failure modes, we compare its predictions against the released gold labels for all 1,452 English test instances. The model produces the correct multi-label vector for 68.3% of cases, which is encouraging—but the remaining 460 errors fall into clear, recurring patterns.

Political over-prediction. The most common error type is flagging political polarization where there is none: 188 of the 919 non-polarized texts (20.5%) receive the political label. In most of these cases, the text *mentions* political figures or insti-

tutions without actually inciting division—think of a newspaper headline about a policy debate rather than a call to action against a political group. The model seems to struggle with the distinction between political *topic* and political *polarization*. This tendency is widespread: the political label co-occurs with the model's predictions at a rate above 72% for every gold category, as if the model treats it almost as a default.

Multi-label under-prediction. Of the 207 instances carrying multiple gold labels, 51.7% are only partially predicted: the model picks up the most obvious polarization type but overlooks the secondary ones. A text that is both racially and religiously polarizing, for example, might receive only the racial label. This pattern suggests the model tends to think in terms of a single dominant category rather than genuinely reasoning about multiple labels at once.

The "Other" trap. Of the 93 times the model predicts "Other," 76 (81.7%) turn out to be false positives—and 49 of those cases are actually political. The category's loose definition makes it an easy fallback, and the model reaches for it too readily. On the opposite end, 13.5% of genuinely polarized texts (72 out of 533) are missed altogether, often because they rely on sarcasm or culturally specific references that lack overt divisive language.

7 Conclusion

This work provides a systematic comparison of fine-tuning, parameter-efficient adaptation, and prompting-based paradigms for polarization type classification under severe label imbalance. Our results demonstrate that, under macro-F1 evaluation, the ability to consistently recover low-frequency labels is a decisive factor. In this setting, a training-free few-shot prompting approach with a large instruction-tuned model proves more effective than both fully fine-tuned and LoRA-based alternatives. While fine-tuned models fail to produce true positives for rare categories—yielding zero F1 for the gender/sexual label—the prompting-based approach successfully captures all label types by leveraging pre-trained semantic knowledge. This advantage translates into strong overall performance, with our system ranking 2nd without any parameter updates.

Future work should explore hybrid approaches that combine the strengths of prompting and supervised learning. In particular, retrieval-augmented example selection, structured prompting strategies such as chain-of-thought reasoning (Wei et al., 2022), and rare-label-aware training objectives represent promising directions for improving both precision and multi-label completeness.

Acknowledgments

This research was supported by the VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91.

Tom Brown and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object

detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Jason Wei and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Prompt Template

The full system prompt appears in Figure 1. The label definitions and classification guidelines are drawn directly from the official *Attitude Polarization Annotation Guideline* published by the task organizers,[4] which defines each polarization type and gives detailed criteria for annotation (e.g., polarization must *clearly* incite division; neutral reporting defaults to non-polarized). We condensed these definitions into concise prompt instructions while keeping their core semantics intact: the strictness criterion, the neutral-text default, and the ambiguity handling rule all come from the original guideline. Each user message begins with the 15 few-shot demonstrations formatted as `Text: "..."` / `Labels: {...}`, followed by the target text.

B Few-Shot Example Selection

Table 4 summarizes the construction of few-shot demonstrations from the English training set. We adopt a 2:1 ratio of neutral to polarized examples, roughly reflecting the underlying label distribution (63.5% non-polarized) and encouraging conservative predictions.

The demonstrations are randomly shuffled and inserted between the system instruction and the target instance. Each example follows the same format as the expected output, enabling the model to infer the JSON structure directly from in-context examples.

[4] https://docs.google.com/document/d/e/2PACX-1vT50PbLDLr_-dHDup-RR2FMIvFEMa1kao6QU0eDELuGTHfU0ZMR7FQLkvLwgWKwFIPHx-fmPjgCA1OH/pub

System message:

You are a neutral and objective classifier for detecting polarization. Your task is to determine whether a given text **clearly** incites division, intolerance, or conflict between groups.

Types of polarization
1. **Political/Ideological** – hostility or conflict between political parties, ideologies, or their followers.
2. **Racial/Ethnic** – hostility or incitement of conflict based on ethnic identity or racial origin.
3. **Religious** – hostility or incitement of conflict between religious groups or followers.
4. **Gender/Sexual** – hostility, discrimination, or marginalization based on gender or sexual orientation.
5. **Other** – hostility targeting other groups or identities (economy, technology, media, etc.).

Classification guidelines
– Assign **1** only if the text clearly promotes division, intolerance, or conflict.
– Neutral reporting, questions, or balanced discussions should receive **0** for all categories.
– When the text is ambiguous or merely mentions a sensitive topic without a polarizing stance, default to **0**.

Output a **JSON** object only.

Figure 1: Complete system prompt. The user message appends 15 few-shot examples and the target text.

Table 4: Few-shot example composition: 1 positive per label + 10 non-polarized. Examples are fixed for reproducibility (random seed 42) and shuffled before insertion into the prompt.

Category	Count	Selection
Political	1	random (seed=42)
Racial/Ethnic	1	random (seed=42)
Religious	1	random (seed=42)
Gender/Sexual	1	random (seed=42)
Other	1	random (seed=42)
Non-polarized	10	random (seed=42)
Total	**15**	