PolAR Bears at SemEval-2026 Task 9: Parameter-Efficient Fine-Tuning and Cross-Lingual Augmentation for Multilingual Polarization Detection

Vinay Babu Ulli
Oogwai Analytics
Bangalore, India
ullivinaybabu@gmail.com

Jyoti Kumari
Department of Linguistics
Banaras Hindu University
jyoti@bhu.ac.in

Abstract

This paper describes our system for SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. We focus on four low-resource Indian languages (Hindi, Bengali, Telugu, and Odia) across three subtasks: Polarization Detection, Type Classification, and Manifestation Identification. To address data scarcity, we employ cross-lingual data augmentation using IndicTrans2, expanding our dataset fourfold. Our unified architecture leverages Qwen3-4B-Instruct optimized via QLoRA, training a linear classification head on masked mean-pooled hidden states. Our system achieved an average Macro F1 of 0.813 across all languages in Subtask 1. For the multi-label frameworks of Subtasks 2 and 3, our results expose a significant pre-training bias within foundational LLMs; while Hindi maintained strong F1 scores of 0.7008 and 0.7248, performance dropped considerably for the other three languages, highlighting ongoing challenges in cross-lingual transfer for nuanced rhetorical techniques.

1 Introduction

The rapid proliferation of digital communication platforms has inadvertently catalyzed the spread of polarized online discourse, exacerbating societal divides along political, racial, religious, and gender lines. Identifying and mitigating such polarizing content is critical, yet deeply challenging due to the linguistic, cultural, and event-specific nuances inherent in global online interactions. To advance research in this domain, SemEval-2026 Task 9 introduces a comprehensive benchmark designed to analyze polarized narratives across diverse contexts. The shared task is divided into three critical components: Multilingual Polarization Detection (Subtask 1), Polarization Type Classification (Subtask 2), and Polarization Manifestation Identification (Subtask 3).

While the shared task encompasses 22 distinct languages, our team's participation strategically focuses on four major Indian languages: Hindi, Bengali, Telugu, and Odia. A persistent challenge in developing robust Natural Language Processing (NLP) models for these languages is the relative scarcity of high-quality, task-specific annotated data. To address the multilingual and low-resource nature of the task, we propose a robust cross-lingual translation-based data augmentation pipeline. Utilizing IndicTrans2 (Gala et al., 2023), a state-of-the-art 1B-parameter Indic-to-Indic neural machine translation model, we systematically translate the original training samples across all four target languages. This translates every instance into a 12-pair translation matrix (each language to every other) while strictly preserving the original polarization labels. This approach expands our training corpus approximately fourfold from 11,350 to 43,244 samples significantly enhancing the model's exposure to diverse linguistic structures.

At the core of our system is the Qwen3-4B-Instruct Large Language Model (LLM). To balance computational efficiency with high task performance, we utilize Parameter-Efficient Fine-Tuning (PEFT) via QLoRA (Dettmers et al., 2024). By leveraging 4-bit NormalFloat (NF4) quantization and applying LoRA adapters across all attention and MLP projection layers, we achieve highly efficient training on an A100 GPU, training only $\sim$33M parameters (a mere 0.81% of the full 4B model).

Across all three subtasks, we maintain a unified architectural and prompting framework. Inputs are structured using Qwen's native ChatML templates, where task-specific system instructions delineate either the polarization labels (Subtask 1), target group categories (Subtask 2), or specific Manifestation Identification (Subtask 3). Instead of autoregressive generation, we formulate the tasks as classification problems by training a linear classi-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2209–2215

fication head on the masked mean-pooled hidden states of the final transformer layer.

Our official CodaBench evaluation results reveal compelling insights into the capabilities and limitations of multilingual LLMs:

- **Subtask 1 (Polarization Detection):** The model, optimized using cross-entropy loss, achieved an average Macro F1 of 0.813. The system demonstrated strong cross-lingual consistency, peaking at 0.8668 for Telugu, indicating that our augmentation successfully aligned foundational semantic spaces for binary detection.
- **Subtasks 2 and 3 (Type and Manifestation Identification):** Formulated as multi-label classification challenges, these tasks achieved average Macro F1 scores of 0.354 and 0.303, respectively. However, a granular analysis reveals a stark disparity: the model performed well in Hindi (Macro F1 of 0.7008 in Subtask 2 and 0.7248 in Subtask 3) but struggled significantly with Bengali, Odia, and Telugu.

Overall, combining base LLMs with cross-lingual data augmentation and PEFT yields an effective approach for core polarization detection. However, our findings highlight that for highly subjective, multi-label rhetorical techniques, foundational LLM pre-training biases and translation artifacts continue to impede equitable performance across lower-resource Indic languages.

2 Related Work

Early NLP research on toxic discourse focused primarily on binary hate speech classification (Waseem and Hovy, 2016; Davidson et al., 2017). Recently, however, work has shifted toward modeling nuanced phenomena like echo chambers and ideological polarization, emphasizing specific rhetorical techniques like dehumanization (Mendelsohn et al., 2021). Addressing the critical need for cross-cultural evaluation, the POLAR benchmark (Naseem et al., 2026b) was introduced. Building on this, SemEval-2026 Task 9 (Naseem et al., 2026a) challenges the community to detect multievent and multicultural online polarization using a highly granular taxonomy.

Multilingual classification is challenging for low-resource languages. While massively multilingual PLMs like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are standard, domain-specific models like IndicBERT (Kakwani et al., 2020) better capture the morphological richness of Indian languages. To overcome data scarcity for Hindi, Bengali, Telugu, and Odia, we employ a cross-lingual augmentation strategy using IndicTrans2 (Gala et al., 2023). Although effective for broad classification tasks, machine translation often struggles to preserve subtle rhetorical nuances, and downstream models frequently exhibit performance biases toward their dominant pre-training languages.

Large Language Models (LLMs) like Qwen (Yang et al., 2025) achieve state-of-the-art results, but full fine-tuning is computationally prohibitive. Parameter-Efficient Fine-Tuning (PEFT), particularly LoRA (Hu et al., 2022) and its quantized variant QLoRA (Dettmers et al., 2024), drastically reduces memory overhead by updating only a small subset of low-rank matrices. Furthermore, while LLMs are optimized for autoregressive generation, generative prompting for complex multi-label classification can suffer from hallucination and thresholding inflexibility. Repurposing decoder-only LLMs as feature encoders, by projecting masked mean-pooled hidden states through a linear head, yields much more stable probabilities (Muennighoff, 2022). We adopt this exact encoder formulation via QLoRA to efficiently handle the multi-label complexity of Subtasks 2 and 3.

3 Dataset Description

The dataset provided for SemEval-2026 Task 9 is uniquely designed to capture online polarization across multievent, multicultural, and multilingual dimensions. While the global dataset spans 22 languages, our experiments specifically filter and target the four prominent Indian languages provided in the task: **Hindi, Bengali, Telugu, and Odia**.

Dataset Statistics and Labels The original unaugmented training subset for our chosen target languages consists of approximately 11,350 text samples. The annotations vary depending on the subtask:

- **Subtask 1:** Binary labels indicating the presence or absence of polarization.
- **Subtask 2 (Target Groups):** Multi-label annotations spanning five categories: *political, racial/ethnic, religious, gender/sexual, and other*.
- **Subtask 3 (Manifestation Identification):** Multi-label annotations spanning six rhetorical

strategies: *stereotype, vilification, dehumanization, extreme language, lack of empathy, and invalidation*.

Augmented Corpus By leveraging IndicTrans2, we overcome the severe data imbalance and low-resource nature inherent to these Indic languages. By generating 12 distinct translation pairs (translating each instance into the other three languages), we scale the training corpus by a factor of roughly $4\times$. Consequently, the final augmented training dataset comprises **43,244 samples**, providing a much denser semantic space for the Qwen3-4B model to align cross-lingual representations.

4 System Description

Our pipeline modifies a decoder-only Large Language Model (LLM) into a parameter-efficient encoder-classifier to handle all three subtasks within a unified framework.

4.1 Task Formulation & Architecture

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent the dataset. We model Subtask 1 as standard classification, and Subtasks 2 and 3 as multi-label classification where $y_i \in \{0, 1\}^K$ ($K = 5$ and $K = 6$, respectively).

We utilize the **Qwen3-4B-Instruct** (Yang et al., 2025) as our base architecture. Inputs are formatted using Qwen's ChatML template, where the system prompt enumerates the subtask categories, and the user prompt contains the raw text x_i. Rather than autoregressive generation, we extract the hidden states of the unmasked text tokens from the final transformer layer. We apply mean-pooling to obtain a dense sequence-level representation, which is then projected through a randomly initialized linear classification head.

4.2 Parameter-Efficient Fine-Tuning (QLoRA)

To maximize memory efficiency, the base model weights are frozen and quantized to 4-bit NormalFloat (NF4) via QLoRA (Dettmers et al., 2024). We inject Low-Rank Adaptation (LoRA) (Hu et al., 2022) modules across all attention and MLP projection layers. Using a rank of $r = 16$, an alpha of $\alpha = 32$, and a dropout of 0.05, we isolate learning to $\sim$33 million trainable parameters (just 0.81% of the model footprint).

4.3 Training and Inference Dynamics

Models were trained on a single A100 GPU utilizing Flash Attention 2 and the AdamW optimizer (learning rate 3e-4, effective batch size 32). For **Subtask 1**, the classification head was optimized using Cross-Entropy (CE) loss for 1 epoch with a 10% linear warmup.

For the multi-label frameworks of **Subtasks 2 and 3**, we optimized using Binary Cross-Entropy (BCE) with logits loss for up to 600 steps (60-step linear warmup). We utilized early stopping, evaluating on the development set every 200 steps based on Macro F1. During inference for Subtasks 2 and 3, we applied a static 0.5 decision threshold post-sigmoid activation to yield the final binarized predictions.

5 Experimental Setup

5.1 Evaluation Metrics

Following the official SemEval-2026 Task 9 evaluation guidelines, the primary evaluation metric for all three subtasks is the **Macro F1-score**. For the multi-label classification tasks (Subtasks 2 and 3), the Macro F1-score is calculated by computing the F1-score independently for each class (target group or technique) and then taking the unweighted average. This ensures that the model's performance is evaluated fairly across all classes, regardless of inherent class imbalances in the training data.

5.2 Hyperparameters and Environment

All experiments and model training were conducted on a single NVIDIA A100 GPU with 40GB of VRAM. To optimize memory utilization during the fine-tuning of the Qwen3-4B-Instruct model, we utilized Flash Attention 2 alongside 4-bit NF4 quantization.

The models were optimized using the AdamW optimizer with a learning rate of 3e-4. We maintained a physical batch size of 16 with a gradient accumulation step of 2, resulting in an effective batch size of 32.

5.3 Training Dynamics

For **Subtask 1**, the model was trained for exactly one epoch, incorporating a linear learning rate warmup over the first 10% of total training steps.

For **Subtasks 2 and 3**, due to the complexity of the multi-label loss landscape, models were trained for up to 600 steps with a fixed linear warmup of 60 steps. To prevent overfitting on the augmented

Hyperparameter	Value
Base Model	Qwen3-4B-Instruct
Precision	4-bit NF4 (QLoRA)
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	All Attention & MLP
Trainable Parameters	$\sim$33M (0.81%)
Optimizer	AdamW
Learning Rate	3e-4
Effective Batch Size	32
Max Epochs (ST 1)	1 Epoch
Max Steps (ST 2 & 3)	600 Steps
Warmup Steps	10% (ST1) / 60 steps (ST2, 3)

Table 1: Summary of hyperparameters and training configurations used across all subtasks.

dataset, we implemented an early stopping mechanism. The models were evaluated on the designated development set every 200 steps, and training was halted if the Macro F1-score failed to improve. At inference, a rigid logit threshold of 0.5 was applied post-sigmoid activation to yield the final multi-label predictions.

To mitigate the risk of overfitting and avoid leaderboard-driven tuning, all hyperparameters and early stopping criteria were tuned strictly on the official development set using a single hold-out strategy. No iterative probing of the test set was conducted.

6 Results and Analysis

6.1 Overall Performance

Our unified Qwen3-4B encoder-classifier demonstrated robust capabilities in **Subtask 1 (Polarization Detection)**, achieving an average Macro F1 of 0.813. As shown in Table 2, the system showcased consistency across the languages for this binary task, peaking with Telugu at 0.8668. This indicates that our augmented semantic spaces successfully allowed the model to distinguish polarizing content across different scripts.

Conversely, performance sharply diverged in the multi-label frameworks of **Subtasks 2 and 3**, yielding average Macro F1 scores of 0.354 and 0.303, respectively. This drop is attributed to the inherent task difficulty and sparse data distribution of multi-label classification over highly overlapping rhetorical categories. Furthermore, these tasks exposed a stark pre-training bias: the model performed well in Hindi ($\sim$0.70–0.72 F1) but struggled significantly with Bengali, Odia, and Telugu.

6.2 Translation Quality Analysis

To validate the central assumption of our cross-lingual augmentation pipeline, that machine translation preserves semantic content and task labels, we performed an automatic evaluation of translation quality using embedding-based similarity.

We computed sentence-level cosine similarity using `bge-m3` embeddings (Chen et al., 2024) for all translated pairs (43,069 distinct sentences). For each source sentence, we measured the similarity between the original text and its translations into the other three target languages. This provides a language-agnostic proxy for semantic preservation across the augmented dataset.

Results: Across all language pairs, mean similarity scores remained above 0.70, indicating generally strong semantic alignment. Telugu and Odia source sentences showed the highest consistency, with similarity ranges of 0.828–0.865 and 0.808–0.817, respectively. Hindi source sentences yielded moderate alignment (0.744–0.774), while Bengali exhibited the lowest consistency (0.703–0.750). Notably, a higher proportion of low-similarity cases (< 0.6) was observed for Bengali-origin translations, particularly in Bengali$\rightarrow$Odia pairs (up to 17.6%), suggesting increased semantic drift in these directions.

Discussion: These findings indicate that IndicTrans2 largely preserves meaning for cross-lingual augmentation, particularly for Telugu and Odia. However, the relatively lower similarity observed for Bengali translations suggests that semantic degradation is more likely for certain source-target combinations. This aligns with our downstream results in Subtasks 2 and 3, where Bengali (and Odia) showed reduced performance. We hypothesize that translation-induced semantic drift disproportionately affects the modeling of fine-grained rhetorical categories, where subtle linguistic cues are critical. Overall, this analysis supports the effectiveness of translation-based augmentation for broad semantic tasks (Subtask 1) while highlighting its limitations for nuanced multi-label classification.

6.3 Language-Specific Variances and Pre-training Bias

The most striking observation from our results is the stark disparity in performance between Hindi and the other three languages in the multi-label tasks (Subtasks 2 and 3). In Subtask 2, Hindi

Task Configuration	Bengali	Hindi	Odia	Telugu	Test Avg F1
Subtask 1: Polarization Detection	0.8184	0.8085	0.7589	0.8668	**0.813**
Subtask 2: Type Classification	0.2347	0.7008	0.3522	0.1299	**0.354**
Subtask 3: Manifestation Identification	0.1570	0.7248	0.1196	0.2102	**0.303**

Table 2: Official CodaBench test results (Macro F1) for our unified Qwen3-4B + IndicTrans2 system.

achieved a strong Macro F1 of 0.7008, while Odia, Bengali, and Telugu dropped to 0.3522, 0.2347, and 0.1299 respectively. A near-identical trend occurred in Subtask 3, where Hindi scored 0.7248, vastly outperforming the others.

We attribute this disparity to two primary factors:

1. **LLM Pre-training Bias:** Although Qwen3-4B is a highly capable multilingual model, its pre-training corpus inherently contains significantly more Hindi text compared to Bengali, Telugu, and Odia. While this discrepancy is masked in the simpler binary classification of Subtask 1, the nuanced reasoning required to identify subtle techniques like *lack of empathy* or *invalidation* (Subtask 3) relies heavily on the model's native language comprehension, heavily favoring Hindi.

2. **Translation Artifacts:** While IndicTrans2 provides state-of-the-art translations, rhetorical nuances that define specific Manifestation Identification (e.g., sarcasm or subtle dehumanization) are notoriously difficult to preserve during machine translation. As a result, the augmented data for Odia, Bengali, and Telugu may have suffered from semantic flattening, making the multi-label boundaries harder for the classification head to resolve.

6.4 Error Analysis and Thresholding Limitations

Beyond language bias, the absolute drops in Subtasks 2 and 3 point to limitations in our inference strategy. At inference, we applied a strict, static 0.5 probability threshold across all logits post-sigmoid activation. In multi-label classification of subjective rhetorical techniques, class distributions are rarely uniform. Overt techniques like *extreme language* naturally yield higher confidence scores from the LLM, whereas implicit biases yield lower logit values. A rigid 0.5 threshold likely resulted in high false-negative rates for the more subtle classes, severely penalizing the Macro F1 score.

Despite these challenges in the granular subtasks, the overall architecture, tuning only $\sim$33M parameters via QLoRA, proved highly memory-efficient and exceedingly effective for core polarization detection (Subtask 1) across heavily under-resourced Indian languages.

7 Conclusion

We presented a unified encoder-classifier system leveraging Qwen3-4B and QLoRA for SemEval-2026 Task 9. To combat data scarcity in Hindi, Bengali, Telugu, and Odia, we expanded the training corpus fourfold using IndicTrans2. Our system proved highly effective for binary polarization detection (Subtask 1), achieving an average Macro F1 of 0.813. However, results on the multi-label Subtasks 2 and 3 exposed critical disparities; the model performed well on Hindi but struggled on the other languages. This highlights that while cross-lingual augmentation and PEFT create robust semantic spaces for broad classification, complex rhetorical identification remains severely bottlenecked by LLM pre-training biases and translation artifacts. Future work must prioritize native dataset curation and class-aware threshold tuning to achieve equitable multilingual performance.

8 Limitations

While our system achieves strong binary classification results, we acknowledge several methodological limitations:

Lack of Baseline Comparisons: Due to computational and time constraints during the evaluation phase, we did not benchmark our system against traditional masked-language models (e.g., mBERT, IndicBERT) or an unaugmented Qwen3-4B baseline. Consequently, while our final system achieved high scores in Subtask 1, the absolute performance gain directly attributable to our cross-lingual augmentation strategy versus the base model's inherent capacity remains unquantified.

Translation Artifacts without Human Evaluation: While our `bge-m3` embedding analysis (Section 6.2) provides a macro-level proxy demonstrating acceptable semantic preservation, we did

not perform manual human evaluation or back-translation checks. Nuanced rhetorical markers (e.g., sarcasm, implicit invalidation) can easily be flattened by machine translation even if dense embedding similarity remains high, limiting the ceiling of our augmented data for Subtasks 2 and 3.

Suboptimal Thresholding: For Subtasks 2 and 3, we applied a static 0.5 logit threshold across all classes. In multi-label classification of subjective techniques, class distributions are rarely uniform. A systematic class-aware threshold sweep would likely reduce the high false-negative rates observed in subtle categories.

Pre-training Bias: The stark performance gap between Hindi ($\sim$0.70 F1) and Bengali/Telugu/Odia ($\sim$0.15–0.35 F1) in Subtasks 2 and 3 highlights a severe limitation in current LLMs. The Qwen base model's pre-training corpus heavily favors Hindi, rendering it naturally more adept at resolving subtle rhetorical boundaries in Hindi compared to other Indic languages.

Acknowledgements

We acknowledge the SemEval-2026 Task 9 organizers for providing the benchmark dataset and evaluation platform. We also extend our gratitude to the AI4Bharat initiative for their open-source IndicTrans2 models, which were instrumental in establishing our cross-lingual data augmentation pipeline.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Jay Gala, Diptesh Kanojia, Sudhanshu Bhatia, Hitesh Doshi, Sumanth Doddapaneni, and 1 others. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions of the Association for Computational Linguistics*, 12.

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

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Pushpak Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Julia Mendelsohn, Vinodkumar Prabhakaran, and Dan Jurafsky. 2021. A framework for the computational linguistic analysis of dehumanization. In *Proceedings of Frontiers in NLP*.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, 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)*. 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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

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

A System Prompts and Reproducibility

To ensure full reproducibility, all cross-lingual augmented datasets, fine-tuned QLoRA adapter weights, and inference scripts are publicly hosted on Hugging Face: https://huggingface.co/collections/vinaybabu/semeval-task9.

We formatted all inputs using the standard ChatML template. The system prompt was used to strictly enumerate the classification categories. Below is an example of the exact prompt structure used for Subtask 3 (Manifestation Identification):

```
<|im_start|>system
You are an expert sociolinguistic AI. Analyze
the following text and identify if it contains
any of the following polarization manifestations:
[Stereotype, Vilification, Dehumanization,
Extreme Language, Lack of Empathy, Invalidation].
Output your analysis based solely on these categories.
<|im_end|>
<|im_start|>user
[INPUT TEXT]
<|im_end|>
```

(Note: Similar enumerating prompts were utilized for Subtask 1 and Subtask 2 based on their respective label sets).