PingAn-NLP at SemEval-2026 Task 9: Multi-Stage Alignment via GRPO and Tiered Ensemble Voting for Multilingual Polarization Detection

DiYang Chen and YouZhen Pang
Ping An Insurance Company of China, Ltd
Shenzhen, China
ytimespace@gmail.com, pangyouzhen@live.com

Abstract

This paper presents the PingAn-NLP system developed for SemEval-2026 Task 9, focusing on multilingual online polarization identification across 18 languages. We propose a multi-stage optimization framework combining Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO). To overcome label imbalance and linguistic nuances, we utilized synthetic reasoning chain augmentation via a high-capacity teacher model (Qwen3-235B) and developed a Smart-Tradeoff reward mechanism to balance precision and recall during reinforcement learning. A language-aware tiered ensemble voting strategy was further implemented to optimize inference performance across diverse linguistic tracks. Our 8B-GRPO-Vote configuration achieved the highest Macro-F1 scores among our experimental variants in 7 out of 18 languages. Officially, our system secured second place in the Bengali, English, Odia, and Turkish tracks.

1 Introduction

The detection of online polarization is critical for understanding social dynamics and preventing the spread of harmful discourse particularly in an era defined by the continuous evolution of mobile technology and the internet.

SemEval-2026 Task 9(Naseem et al., 2026a) presents a significant challenge by requiring fine-grained hate speech detection across 18 languages(Naseem et al., 2026b). The primary obstacles include severe data imbalance, the subtle nature of multicultural expressions, and the scarcity of positive samples in low-resource languages.

Our leverages the multilingual capabilities of the Qwen series (Qwen-32B and Qwen-8B)(Yang et al., 2025). We bridge the semantic gap using a teacher-student distillation method where reasoning chains are generated to enrich the training signal. Furthermore, we optimize the model's alignment using GRPO(Shao et al., 2024), specifically designing a reward function to enhance recall in minority classes while preventing mode collapse.

2 Background

Polarization research extends beyond binary hate speech detection to capture rhetorical strategies such as stereotyping, vilification, dehumanization, and empathy erosion. The POLAR benchmark(Naseem et al., 2026b)frames this as a multilingual, multi-label task spanning culturally diverse and low-resource settings. Unlike explicit toxicity, polarization is often implicit and context-dependent, posing challenges under severe label imbalance.

Recent alignment methods, including Reinforcement Learning from Human Feedback(Christiano et al., 2017), aim to improve reasoning stability in LLMs. Group Relative Policy Optimization (GRPO) enables efficient policy refinement through relative comparisons among sampled outputs.

Building on these foundations, we formulate multilingual polarization detection as a reasoning-aligned optimization problem. We combine synthetic reasoning distillation with task-aware GRPO reward design to stabilize cross-lingual decision boundaries while addressing recall–precision tradeoffs in highly imbalanced settings.

3 System Overview

3.1 Synthetic Reasoning Chain Augmentation

To provide the student models (Qwen3-8B/32B) with deeper logic, we employed a high-capacity teacher model (Qwen3-235B) to generate synthetic reasoning chains for the training set. This process, known as reasoning augmentation, distills the complex cross-lingual logic of the teacher

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2810–2816

model into intermediate semantic signals, helping the student model understand the "why" behind each label. The specific prompt template used for generating these reasoning chains is detailed in **Appendix B**.

3.2 Reward Function Design

A critical challenge in this multi-label task is the extreme label imbalance and the tendency of models to collapse into all-zero predictions. To address this, we developed a composite reward mechanism for GRPO, consisting of a *Semantic Accuracy Reward* and a *Smart-Tradeoff Reward*.

3.2.1 Semantic Accuracy Reward (R_{acc})

The primary goal of R_{acc} is to guide the model toward precise label alignment. We combine Macro-F1 score and Hamming Accuracy to provide a smooth gradient:

$$R_{acc} = 0.4 \cdot \text{Hamming}(\hat{y}, y) + 0.6 \cdot \text{F1}_{macro}(\hat{y}, y) \quad (1)$$

Where $\hat{y}$ and y are the predicted and ground-truth label vectors. This function balances the "all-or-nothing" nature of F1 scores with Hamming Accuracy, which rewards the model for each correctly identified individual label, even if the overall multi-label set is not perfect. We specifically set `zero_division=1` for F1 calculation to properly reward the model for correctly predicting "all-zero" vectors in non-polarized samples.

3.2.2 Smart-Tradeoff Reward (R_{trade})

To specifically combat the high false-negative rate in low-resource languages like Hausa, we implemented a conditional logic based on the presence of positive labels:

- **Positive Samples ($\sum y > 0$):** We prioritize Recall. If the model fails to detect any polarization (all-zero prediction), it receives a heavy penalty of -1.0. If it detects at least one correct label, the reward is calculated as $0.2 + 0.8 \cdot \text{Recall}$.
- **Negative Samples ($\sum y = 0$):** We prioritize Precision. An all-zero prediction (correct negative) is rewarded with +0.5, while any false-positive detection is penalized with -0.5.

This "smart-tradeoff" logic acts as a dynamic training signal. It forces the model to be extremely sensitive when polarization is present (addressing the "missing" hate speech problem) while rewarding caution when the input is clean, preventing the model from hallucinating labels.

3.2.3 Format and Penalty Constraints

Consistent with the formatting requirements of the task, any generation that fails to follow the specified output format or results in an unparseable completion is assigned a penalty of -0.5. This ensures that the model maintains high structural integrity during the reinforcement learning phase.

3.3 Tiered Ensemble Voting Strategy

To mitigate the inherent variance in LLM generations and optimize the precision-recall tradeoff across diverse linguistic contexts, we implemented a tiered ensemble voting strategy during the inference phase.

Diversity Generation For each test instance, we utilize the `vLLM` library to perform parallel decoding, generating $N = 5$ independent samples with a temperature of $T = 0.8$. This high-temperature setting ensures sufficient semantic diversity among candidates, which is essential for the subsequent voting process.

Language-Specific Thresholding Recognizing that 18 languages exhibit varying degrees of data scarcity and label difficulty, we categorize them into three tiers with distinct decision thresholds:

- **Liberal Strategy ($V \geq 1$):** Applied to low-resource languages such as Hausa (HA) and Odia (OR). In this tier, a label is assigned if at least one out of five samples predicts it as positive. This serves as a "recall booster" for languages where polarization is often under-reported by single-pass inference.
- **Stable Strategy ($V \geq 3$):** Applied to high-resource languages like Hindi (HI) and Arabic (AR). A majority-vote (3/5) requirement is used here to prioritize precision and filter out noise in well-represented tracks.
- **Moderate Strategy ($V \geq 2$):** Applied to intermediate languages like English (EN) and German (DE). A 2/5 threshold is used to strike a balance between detection sensitivity and robustness.

Instead of relying on a single "lucky" guess from the model, we ask it the same question five

times in slightly different ways. We then decide the final answer based on the language's difficulty. For very hard languages where the model is "shy" about labeling hate speech, we are more liberal: if the model sees it even once, we count it. For easier languages, we are more strict and require a consensus among the majority of its answers.

3.4 Mathematical Context and Rationales

To provide a transparent view of our optimization process, we detail the mathematical framework of Group Relative Policy Optimization (GRPO) and provide intuitive explanations for our design choices as encouraged by the task organizers (Naseem et al., 2026c).

Group-Based Advantage Estimation In GRPO, for each input prompt, the model generates a group of G outputs $\{o_1, o_2, ..., o_G\}$. The advantage A_i for each output is calculated relative to the group's performance:

$$A_i = \frac{R(o_i) - \text{mean}(R)}{\text{std}(R) + \epsilon} \quad (2)$$

Instead of comparing a model's answer to a fixed "correct" score, GRPO makes the model compete against itself within a small group. If one generation is better than the average of its peers, it receives a positive signal; if it is worse, it is penalized. This "relative" scoring allows the model to refine its nuances in multi-label classification without requiring a massive, separate critic model, significantly saving computational resources on our NPU cluster.

4 Experimental Setup

To evaluate the effectiveness of various modeling strategies, we conducted experiments across a spectrum of configurations, ranging from zero-shot baselines to specialized reinforcement learning-based models.The specific prompt template utilized for our zero-shot inference is provided in **Appendix A**.

4.1 Model Configurations

We define the primary configurations utilized in our evaluation as follows:

- **Qwen-235B**: Represents the zero-shot inference performance of the Qwen3-235B model, serving as a high-capacity linguistic baseline.
- **Gemini**: Refers to zero-shot inference conducted using the Gemini 3 Flash model.
- **8B-Full-SFT**: Denotes the Qwen3-8B model after undergoing full-parameter supervised fine-tuning (SFT) on the comprehensive task dataset.
- **32B-Lora-SFT**: Represents the Qwen3-32B model fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021)to balance computational efficiency and model capacity.
- **32B-SFT-Vote**: An ensemble approach where a voting strategy is applied to multiple outputs generated during the inference phase of the 32B-Lora-SFT model.
- **8B-GRPO**: A specialized version of the 8B-Full-SFT model that underwent further optimization using Group Relative Policy Optimization (GRPO) to refine its classification boundaries for difficult language subsets.
- **8B-GRPO-Vote**: Represents our advanced ensemble configuration where a voting strategy is applied to the outputs of the 8B-GRPO model. This step is designed to consolidate the specialized reasoning capabilities developed during the reinforcement learning phase and mitigate individual generation variances.

4.2 Training and Implementation Details

All fine-tuned models were implemented using the Hugging Face Transformers(Wolf et al., 2020) and TRL libraries. For **8B-Full-SFT**, we utilized a learning rate of 1×10^{-6} with DeepSpeed ZeRO-2 optimization(Rasley et al., 2020). The **32B-Lora-SFT** was trained with a rank (r) of 64 and alpha (α) of 128.

For the **GRPO** phase, we employed a group size of 16 generations per prompt and a KL-divergence coefficient (β) of 0.04 to prevent mode collapse. Training was performed on a distributed NPU cluster, ensuring efficient scaling for both the 8B and 32B architectures.

5 Results and Analysis

Our experimental results, as detailed in Table 1, reveal several critical insights into the performance of different model architectures and training strategies across 18 linguistically diverse tracks.

Table 1: Comprehensive comparison of Macro-F1 scores across 18 languages. Our "8B-GRPO-Vote" system demonstrates competitive performance, particularly in high-resource and specialized language contexts.

Language	Qwen-235B	Gemini	8B-Full-SFT	32B-Lora-SFT	8B-GRPO	32B-SFT-Vote	8B-GRPO-Vote
Amharic	0.3207	**0.5325**	0.3946	0.4095	0.4115	0.4254	0.4450
Arabic	0.4916	0.5315	0.5463	0.5559	0.5658	0.5500	**0.6028**
Bengali	0.2034	0.2228	0.1810	0.1938	0.1526	**0.2554**	0.2058
German	0.3321	0.3592	0.4336	0.4500	0.4643	0.4664	**0.4938**
English	0.3407	0.4079	0.4333	0.4331	0.4609	0.4766	**0.5071**
Persian	0.2686	**0.4249**	0.2832	0.2727	0.2641	0.3411	0.3246
Hausa	0.0677	0.0821	**0.1188**	0.0625	0.0122	0.0791	0.0491
Hindi	0.3708	0.4526	0.6814	0.6753	0.6767	0.6852	**0.7197**
Khmer	0.0768	0.1037	0.2474	0.2370	0.2248	**0.3273**	0.2891
Nepali	0.5077	0.5581	0.5712	0.5519	0.5649	**0.6293**	0.5705
Odia	0.2069	0.2242	0.2085	**0.3280**	0.1705	0.2489	0.2408
Punjabi	0.2995	0.3919	0.4203	0.4085	0.4288	0.4634	**0.4963**
Spanish	0.3712	0.4230	0.3718	0.3885	0.4290	0.4173	**0.4656**
Swahili	0.2991	0.3769	0.4351	0.4115	0.3965	0.4381	**0.4660**
Telugu	0.1340	0.0721	0.2494	0.2673	0.2763	**0.3517**	0.3102
Turkish	0.4531	**0.5374**	0.4546	0.4646	0.4813	0.4915	0.4957
Urdu	0.4008	0.5992	**0.8023**	0.7894	0.7879	0.7899	0.7979
Chinese	0.5378	0.5051	0.5418	0.6057	0.5264	**0.6453**	0.5622

5.1 Overall Trends and Best Systems

The results demonstrate that our proposed 8B-GRPO-Vote configuration achieved the highest Macro-F1 scores in 7 out of 18 languages, particularly excelling in high-resource and complex reasoning tracks such as English (0.5071), Hindi (0.7197), and Arabic (0.6028). This confirms that combining reinforcement learning via Group Relative Policy Optimization (GRPO) with ensemble voting provides a robust framework for capturing subtle polarization cues.

5.2 Ablation: The Impact of GRPO and Voting

As encouraged by the task organizers to share ablation studies, we observe a clear performance trajectory from SFT to GRPO-enhanced models.

- **From SFT to GRPO**: Comparing **8B-Full-SFT** to **8B-GRPO**, we noticed that while raw SFT provides a strong foundation, the GRPO phase significantly refines the model's decision boundaries. In languages like Spanish and Punjabi, GRPO alone improved the F1 score by approximately 2-5
- **The Voting Advantage**: The transition from **8B-GRPO** to **8B-GRPO-Vote** consistently resulted in performance gains. This suggests that the "wisdom of the crowd" effectively mitigates the high variance often associated with reinforcement learning generations, leading to more stable multi-label predictions.

5.3 Zero-Shot Baselines vs. Supervised Specialists

A notable finding is the performance of zero-shot baselines in specific linguistic contexts:

- **Gemini Dominance**: In tracks such as Amharic (0.5325) and Turkish (0.5374), the zero-shot Gemini model outperformed all fine-tuned variants. This represents a "negative result" for our fine-tuning pipeline in these specific languages, likely due to the base model's superior pre-training coverage or our fine-tuning dataset's failure to capture unique cultural nuances in those regions.
- **SFT for Specific Dialects**: Conversely, in Urdu, the 8B-Full-SFT model reached a peak F1 of 0.8023, vastly outperforming the Qwen-235B zero-shot baseline (0.4008). This indicates that for certain languages, task-specific alignment is far more critical than raw model scale.

5.4 Error Analysis and Negative Insights

Following the requirement for deep insight beyond overall scores, we conducted an error analysis on low-performing tracks:

- **Low-Resource Challenges (Hausa and Odia)**: Despite specialized training, performance in Hausa remained relatively low. We

observed that the model frequently generated "False Negatives" in categories like *Dehumanization*, often failing to recognize localized slurs not present in the teacher model's synthetic reasoning chains.

- **The "Mode Collapse" Risk**: During early GRPO experiments, we encountered a struggle where the model's output standard deviation dropped below 0.05, leading to repetitive, non-informative labels. Increasing the KL-divergence penalty (β) and utilizing the 8B-GRPO-Vote strategy were essential to restoring prediction diversity.

6 Conclusion

In this work, we presented the PingAn-NLP system for SemEval-2026 Task 9 Subtask 3 on multilingual polarization manifestation identification. We proposed a multi-stage optimization framework that integrates supervised fine-tuning (SFT), synthetic reasoning distillation from high-capacity teacher models, and reinforcement learning via Group Relative Policy Optimization (GRPO), complemented by a language-aware ensemble voting strategy. this study suggests that in complex multilingual classification tasks, model scale alone is insufficient. Instead, structured reasoning supervision, reward-driven alignment, and adaptive inference strategies jointly form a robust paradigm for long-tail multilingual polarization detection. Future work will focus on improving culturally grounded reasoning generation, incorporating real human feedback into the reward model, and exploring more adaptive expert routing mechanisms for low-resource languages.

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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. *Preprint*, arXiv:2106.09685.

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, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 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, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

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, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026c. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

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. *Preprint*, arXiv:2402.03300.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

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. *Preprint*, arXiv:2505.09388.

A Verbatim Zero-Shot Prompt Template

The following prompt template was utilized for the zero-shot baseline experiments. It integrates professional persona assignment, multi-dimensional label definitions, and structured output constraints.

System Prompt
You are an expert computational sociolinguist specializing in Online Polarization Detection. Your task is to analyze social media text for specific manifestations of polarization.

User Prompt
Task Description:
Analyze the following text (which may be in a low-resource language like Odia or Khmer) and classify it according to the 6 manifestations of polarization defined below.

Input Text:
"""{input_text}"""

Label Definitions:

- **stereotype**: Generalized beliefs about a group of people that oversimplify their characteristics. Does the message generalize certain characteristics of individuals to all members of a group? Does it ignore individual differences? Stereotypes often reduce complex personalities to simple, one-size-fits-all representations. Hypothetical examples: Migrants/whites/blacks are greedy, cowardly, and sexist. Men are strong. Women are weak.
- **vilification**: Using abusive or insulting language to portray a group as evil, criminal, or dangerous. Does the text defame or demonize a particular group, person, or entity, inciting fear? This could be through exaggeration, misrepresentation, or biased framing that presents the subject in a negative, harmful light. Hypothetical examples: Migrants/whites/blacks are oppressive; a traitor; a bandit.
- **dehumanization**: Denying the humanness of a group, comparing them to animals, insects, or objects. Is the text stripping a group or individual of their human qualities or personality? Dehumanization can be seen in language that compares people to animals, machines, or objects, or that otherwise denies their humanity, dignity, or individuality. Hypothetical examples: Using terms such as hyena, snake, or cockroach to describe persons/groups.
- **extreme_language**: Using hyperbolic, inflammatory, or violent rhetoric. Does the text use extreme language or make absolute statements? Attitude polarization often involves absolutist language (e.g., "always", "never"), extreme (e.g., "worst", "best"), or dichotomous (e.g., "us vs. them", "right vs. wrong"). Hypothetical example: Migrants/Arabs/Russians are never to be trusted. We and they cannot live together.
- **lack_of_empathy**: Showing indifference, mockery, or callousness towards the suffering of others. Does the text lack empathy or understanding for other perspectives or experiences? This could involve marginalizing the viewpoints or experiences of others or showing a lack of willingness to understand or empathize with them. Hypothetical example: Wearing the hijab/cross-dressing is a sign of extremism.
- **invalidation**: Does the text deny or invalidate the identity and existence of people? It involves rejecting the identity and existence of other people or groups. Hypothetical example: There is no nation called Palestine/Israel. We do not allow them to exist.

Chain-of-Thought Reasoning Requirements:
1. Translate & Contextualize: First, translate the text into English (if it is not already) and explain any cultural nuances or metaphors.
2. Step-by-Step Verification: For EACH of the 6 labels, explicitly check if the text matches the definition. Provide evidence from the translated text.
3. Logical Consistency: Ensure the labels are consistent (e.g., Dehumanization often implies Vilification).

Output Format:
Provide the output strictly in JSON format with two keys: "reasoning" (string) and "labels" (object with 0/1 integers).

B Training Data Generation Prompt (Reasoning Distillation)

The following template was used to generate reasoning chains for the supervised fine-tuning and GRPO phases. It directs a high-capacity teacher model to provide logical justifications for existing Ground Truth labels.

System Prompt
You are an expert computational sociolinguist specializing in Online Polarization Detection. Your task is to generate a rigorous "Chain-of-Thought" reasoning process that logically explains why a given text matches its Ground Truth labels.

User Prompt
Input Text: """{text}"""
Ground Truth Labels (1=Present, 0=Absent):
stereotype: {stereotype}; vilification: {vilification}; dehumanization: {dehumanization}; extreme_language: {extreme_language}; lack_of_empathy: {lack_of_empathy}; invalidation: {invalidation}

Task Requirements:
1. **Reverse Engineer the Logic:** You must accept the Ground Truth labels as absolute truth. Your goal is to explain why the text fits these labels based on their academic definitions.
2. **Step-by-Step Analysis:**
– If a label is 1, quote the specific words in the text that trigger this label and explain the link to the definition.
– If a label is 0 (and the decision is non-obvious), briefly explain why the text does not meet the threshold.
3. **Format:** Output a JSON object containing a "thought_process" field (the reasoning) and the "labels" field.

Definitions for Reference:
(The prompt includes the full definitions for stereotype, vilification, dehumanization, extreme_language, lack_of_empathy, and invalidation as detailed in Appendix A.)

Output Format Example:
{
"thought_process": "The text refers to the group as 'rats', which is a classic animalistic metaphor, justifying the [Dehumanization] label. The phrase 'wipe them out' constitutes a call for violence, fitting [Extreme Language]...",
"labels": { "stereotype": 0, "vilification": 1, ... }
}