CYUT at SemEval-2026 Task 9: Monolingual vs. Multilingual LoRA Tuning for Multicultural and Multievent Polarization Detection

Shih-Hung Wu*, **YUN-KUANG LIAO, Shih-Siang Su, Yi-Min Jian**
Department of Computer Science and Information Engineering
Chaoyang University of Technology Wufeng, Taichung, Taiwan
shwu@cyut.edu.tw, s11427613@gm.cyut.edu.tw,
s11427606@gm.cyut.edu.tw, s11427601@gm.cyut.edu.tw

Abstract

This study addresses SemEval-2026 Task 9 on Detecting Multilingual, Multicultural, and Multievent Online Polarization, exploring the performance differences between monolingual and multilingual LoRA (Low-Rank Adaptation) fine-tuning techniques when processing online polarization phenomena. The research points out that online polarization is not only a language phenomenon, but a complex social language problem highly influenced by cultural contexts and event backgrounds. To address the limitation of existing research that only treats polarization as a binary classification, this study participates in three levels of subtasks: Subtask 1: Polarization Detection, Subtask 2: Polarization Type Classification (e.g., politics, religion), and Subtask 3: Manifestation Identification (analyzing rhetorical strategies that construct polarization, such as stereotypes and dehumanization narratives). This study aims to establish a more contextually grounded and diagnostic model analysis framework to enhance the model's generalization ability and fairness in cross-lingual environments. By exploring different fine-tuning configurations to build a robust ensemble system, the experimental results show that our approach demonstrates exceptional proficiency in the Chinese domain, securing the 1st place ranking in Subtask 1 (Polarization Detection) for Chinese. Furthermore, we observe that while the monolingual LoRA strategy exhibits strong performance in specific languages like Chinese, integrating it with multilingual LoRA models via ensembling provides the diverse features crucial for identifying complex cross-cultural rhetoric.

1 Introduction

In contemporary online public discourse, polarization is a pervasive, complex sociolinguistic phenomenon constructed through diverse rhetorical strategies—such as collective stigmatization and exclusionary narratives—across specific cultural and event contexts (Naseem et al., 2026a). Its manifestations and social orientations vary significantly; for instance, election discourse revolves around political stances, while ethnic conflicts often employ dehumanization or intergroup antagonism. Furthermore, cultural norms dictate whether conflict is expressed directly or through subtle rhetorical devices, necessitating strong contextual grounding for accurate analysis. Consequently, treating polarization merely as a binary classification falls short of revealing its diverse social orientations and linguistic constructs. Authentic discourse often simultaneously targets multiple entities through varied rhetoric. Without more granular levels of analysis, cross-cultural comparisons and concrete diagnostics for model error patterns remain unattainable (Vidgen and Derczynski, 2020).

To address this research need, SemEval-2026 Task 9 proposes a tiered, multilingual benchmark that decomposes polarization analysis into three interconnected subtasks. Subtask 1 provides foundational binary identification of polarization. To elucidate deeper social implications, Subtask 2 identifies specific target orientations (e.g., political or ethnic groups), enabling the analysis of systematic biases across contexts. Furthermore, Subtask 3 captures the underlying rhetorical constructs (e.g., stereotypes, dehumanization narratives), linking model predictions with concrete linguistic evidence. By combining multilingual data with this hierarchical structure, Task 9 transcends aggregate performance metrics. It provides an operationalizable analytical platform that supports granular cross-lingual error analysis, model bias diagnostics, and event-level comparisons across diverse linguistic cultures.

2 Related work

Polarized language exacerbates societal fragmentation through dichotomous narratives, boundary

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1621–1631

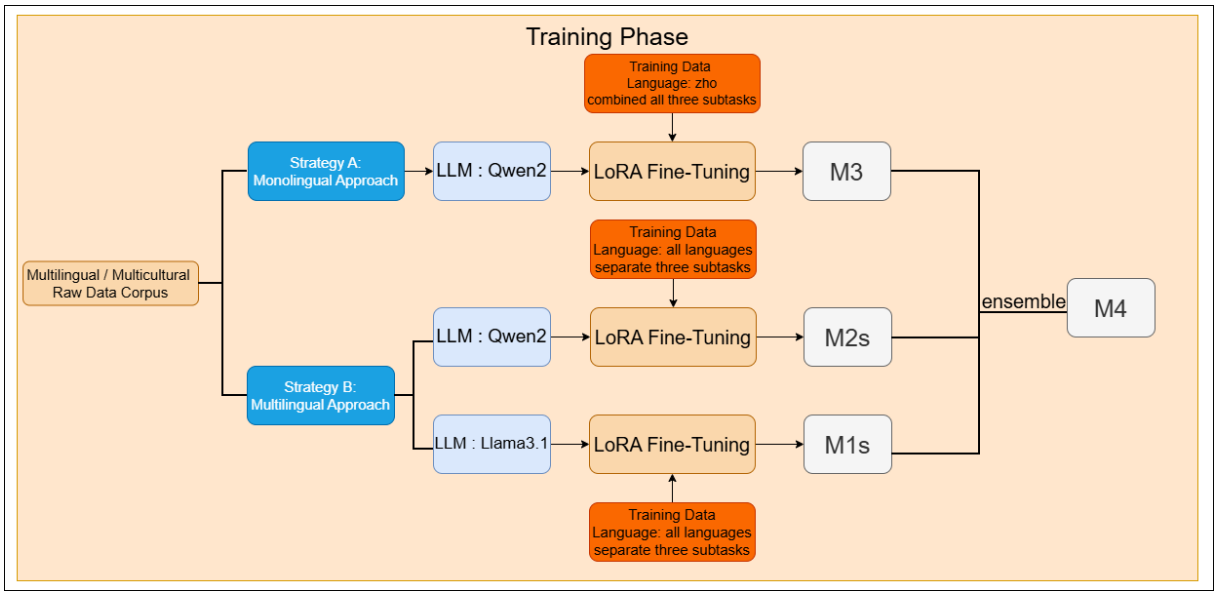

Figure 1: System Framework

ID	Model	Fine-tuning Strategy	Training Data / Logic
M1	Llama-3.1-8B-Instruct	Task-Specific (3.3.1)	Multilingual data, separate tasks
M2	Qwen2-7B-Instruct	Task-Specific(3.3.1)	Multilingual data, separate tasks
M3	Qwen2-7B-Instruct	Task-Fused(3.3.2)	Chinese only, combined tasks
M4	ensemble(M1+M2+M3)	Weighted Ensemble(3.4)	Performance-based weighting

Table 1: M1, M2, and M3 serve as the core models, while M4 represents the ensemble of M1, M2, and M3.

delineation, and implicit normative evaluations, making it far more strategic and context-dependent than explicit hate speech (Vidgen and Derczynski, 2020; Van Dijk, 1998; Wodak, 2015). Consequently, coarse-grained harmful speech detection is insufficient to capture its complexity (Vidgen et al., 2021). Addressing this, SemEval-2026 Task 9 proposes a benchmark that decomposes polarization analysis into three hierarchically distinct subtasks—Polarization Detection, Polarization Type Classification, and Manifestations—echoing the need to analyze targeted social orientations and pragmatic strategies to evaluate social impact (Naseem et al., 2026a).

Linguistic mechanisms like dehumanization and stereotypes propel polarization by compressing diverse stances into in-group versus out-group dichotomies, effectively reducing empathetic engagement (Haslam, 2006; Wodak, 2015). Because these manifestations exhibit strong event-specific dependency, automated identification requires fine-grained language understanding (Vidgen et al., 2021). Furthermore, cross-cultural contexts elevate detection difficulty; different linguistic communities employ vastly different discursive strategies (Ribeiro et al., 2020). Low-resource languages face additional hurdles due to insufficient data and the difficulty of transferring cultural metaphors from high-resource languages (Joshi et al., 2020). The 22 languages covered by Task 9 thus provide a critical benchmark for cross-lingual evaluation(Naseem et al., 2026b).

At the model level, while large pre-trained language models possess robust shared semantic representations, their performance heavily relies on training data configurations (Pfeiffer et al., 2021). Although multilingual joint training can benefit low-resource languages, pragmatic differences may lead to negative transfer in fine-grained tasks (Chen et al., 2024). Consequently, Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA are ideal for exploring cross-lingual generalization (Hu et al., 2022).

Building upon this, we address the complexities of cross-lingual polarization detection by utilizing Llama-3.1-8B-Instruct and Qwen2-7B-Instruct with LoRA fine-tuning. We construct diverse model configurations, encompassing both Chinese monolingual and multilingual joint models across the three subtasks. Rather than conducting a strictly

ID	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
M1	0.681	**0.8415**	0.8119	0.7423	0.8132	0.8088	0.7156	**0.8281**	0.6464	0.568	0.8391
M2	0.5055	0.8318	0.8217	0.7389	0.8019	0.7006	0.698	0.7642	0.5888	0.5303	0.831
M3	0.2488	0.7745	0.7633	0.699	0.7386	0.456	0.5046	0.6395	**0.7432**	0.1132	0.3398
M4	0.6891	0.8399	0.8205	**0.7465**	**0.8153**	0.7705	0.6993	0.8111	0.6674	0.5747	**0.8394**
Baseline	**0.7151**	0.7957	**0.8528**	0.6714	0.7802	**0.8424**	**0.7753**	0.7379	0.6773	**0.6592**	0.821

Table 2: Macro F1 scores of the four core models on the Subtask 1 test set.

ID	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
M1	**0.8837**	0.7102	0.7872	**0.8148**	0.8134	**0.7799**	0.7775	**0.8286**	0.8015	0.7866	0.9258
M2	0.8604	**0.7529**	0.7148	0.8021	0.806	0.7787	**0.7825**	0.8226	0.7812	0.7366	0.9185
M3	0.6562	0.5907	0.4845	0.7176	0.6098	0.6646	0.5443	0.5587	0.746	0.5853	0.9248
M4	0.8712	0.6897	0.722	0.8058	**0.8138**	0.779	0.775	0.827	**0.8055**	0.7758	**0.9315**
Baseline	0.8798	0.7765	**0.7898**	0.7241	0.7457	0.7266	0.7571	0.644	0.6957	**0.789**	0.8691

Table 3: Macro F1 scores of the four core models on the Subtask 1 test set.

controlled variable analysis, our primary objective is to leverage the structural and strategic differences of these models to build a highly robust weighted ensemble system, aiming to maximize detection performance, particularly in the Chinese domain.

3 Methodology

3.1 Experimental Design and Model Architecture

Addressing the multilingual Polarization Detection challenge defined by SemEval 2026 Task 9, this study aims to evaluate the efficacy of different model architectures in capturing nuanced semantic orientations. The experiment selects two representative open-source Large Language Models (LLMs) as the foundation: Qwen2-7B-Instruct and Llama-3.1-8B-Instruct. These two models represent current high-performance multilingual instruction-tuned architectures. This study treats them as independent experimental pipelines to mitigate framework-specific confounding factors, thereby analyzing the differences in polarization feature extraction attributed to the base model architectures and fine-tuning strategies.

3.2 Low-Rank Adaptation (LoRA) and Fine-Tuning Strategy

To achieve Parameter-Efficient Fine-Tuning (PEFT) under constrained computational budgets, this study employs the LoRA technique. By integrating low-rank matrices into the attention mechanisms of the Transformer layers, this study is able to maintain the pre-trained model's generalization ability while tailoring the model to the domain-specific knowledge of Polarization Detection.

3.3 Experimental Model Configurations

Based on the experimental results on the development set in Appendix A, and targeting the data distribution and task attributes, this study designed two comparative data configuration schemes, and constructed the following three core model configurations based on the subsequent fine-tuning strategies for comparative analysis. Table 1 summarizes the four model configurations (M1 to M4) implemented in this study. To comprehensively evaluate the effectiveness of different base models and fine-tuning strategies, we designed three single models (M1-M3) and one ensemble model (M4). It is important to note that M1, M2, and M3 differ across multiple dimensions (base architecture, language data, and task integration) simultaneously; this deliberate variance aims to maximize feature diversity for the subsequent ensemble process rather than serving as strictly controlled variables for causal inference.

Regarding the single models, M1 and M2 aim to compare the performance of different large language models in cross-lingual environments. M1 utilizes Llama-3.1-8B-Instruct, while M2 utilizes Qwen2-7B-Instruct; both are independently fine-tuned using the cross-lingual task-specific fine-tuning strategy. In contrast, M3 is also based on Qwen2-7B-Instruct but adopts a monolingual multi-task joint fine-tuning strategy, training exclusively on pure Chinese data, thereby exploring the potential advantages of single-language and joint-task training.

To optimize feature extraction for these instruction-tuned models, highly customized

ID	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
M1	0.319	0.5257	0.1456	0.4168	0.4644	0.4519	0.0904	0.5693	0.1393	0.2344	0.2908
M2	0.353	**0.5436**	0.23	0.4598	0.4747	0.4903	0.0931	0.6944	0.24	**0.5651**	0.3712
M3	0.021	0.481	0.2416	**0.4955**	**0.4804**	0.2955	0.1099	0.3508	**0.5018**	0.0658	0.0651
M4	**0.4142**	0.536	0.2699	0.4717	0.479	**0.5228**	0.1312	0.7028	0.4836	0.5651	0.3998
Baseline	0.3716	0.4855	**0.2887**	0.4078	0.3333	0.4626	**0.2038**	**0.7911**	0.3759	0.6268	**0.4772**

Table 4: Macro F1 scores of the four core models on the Subtask 2 test set.

ID	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
M1	0.5654	0.1334	**0.4206**	0.457	0.4109	0.4912	0.3364	0.2783	0.5467	0.675	0.6381
M2	0.7071	0.3349	0.3372	0.4567	0.5213	0.5634	0.3309	0.1574	0.5091	**0.7633**	0.7892
M3	0.4484	0.2533	0.1361	0.4706	0.472	0.4818	0.2347	0.1276	0.4985	0.2306	**0.8262**
M4	0.6844	0.3493	0.4117	**0.5061**	0.5903	0.5545	0.3807	0.2503	**0.5875**	0.7183	0.8245
Baseline	**0.7219**	**0.56**	0.365	0.4491	**0.5904**	**0.5935**	**0.4417**	**0.3145**	0.4708	0.7127	0.6697

Table 5: Macro F1 scores of the four core models on the Subtask 2 test set.

prompts were designed (detailed in Appendix B).

Finally, to further enhance the stability and accuracy of the overall predictions, we proposed the ensemble model M4. M4 employs a weighted ensembling strategy, assigning different weights based on the respective performance (F1-Macro scores) of M1, M2, and M3 to aggregate the predictive outputs of the three models.

3.3.1 Cross-Lingual Task-Specific Fine-Tuning

Adopting a "task-decoupled" strategy, independent models are trained separately for the three subtasks. Each model integrates all language data under its respective task. The core motivation is to enhance the models' shared semantic representations of a single task across different linguistic manifestations, thereby reducing cross-task target interference.

3.3.2 Monolingual Multi-Task Joint Fine-Tuning

Adopting a "task-integrated" strategy, this approach utilizes exclusively Chinese language data but integrates the labels of the three subtasks for unified fine-tuning. This scheme aims to evaluate the model's ability to learn cross-task logical dependencies within a monolingual environment, and to further verify the transfer efficacy of this knowledge in cross-lingual scenarios.

3.4 Performance-Oriented Weighted Ensembling Strategy

To further optimize the stability of system outputs, this study developed a performance-based weighted ensembling mechanism. This mechanism dynamically determines weights based on the predictive outputs of the three model configurations (M1, M2, M3), and through weight normalization, ensures the sum of weights equals 1, realizing multi-model decision-level fusion.

3.4.1 Subtask 1 (Binary Classification)

The F1-Macro scores of the three model configurations (M1, M2, M3) across various languages in Subtask 1 serve as the foundational weights.

3.4.2 Subtasks 2 and 3 (Multi-Label Classification)

A more fine-grained label-level weighting is adopted. The F1-Macro performance of each independent label within the three model configurations (M1, M2, M3) is extracted separately, and normalized weights are calculated for each specific label, thereby capturing the discriminative superiority of different models on specific polarization dimensions.

4 Results

Tables 2 to 3 present the results of the four core models of this study across various languages in Subtask 1.

Tables 4 to 5 present the results of the four core models of this study across various languages in Subtask 2.

Tables 6 to 7 present the results of the four core models of this study across various languages in Subtask 3.

From these tables, the performance under different models and different fine-tuning strategies can be observed. In Subtask 1 and Subtask 3, the aggregate performance of M1 is the best, while in Subtask 2, the aggregate performance of M2 is

ID	amh	arb	ben	deu	eng	fas	hau	hin	khm
M1	0.3044	**0.5735**	0.0805	**0.4555**	**0.4916**	**0.3763**	0.0762	0.6682	0.1322
M2	0.1958	0.4869	0.0792	0.3211	0.2608	0.1374	0	**0.7184**	0.2163
M3	0.0181	0.3261	**0.1386**	0.2695	0.3232	0.1687	0.0273	0.2759	0.0238
M4	0.3677	0.5594	0.1198	0.401	0.4222	0.3742	0.0828	0.6947	0.2248
Baseline	**0.4433**	0.3902	0.0868	0.3485	0.41	0.2004	**0.7456**	0.2348	**0.6095**

Table 6: Macro F1 scores of the four core models on the Subtask 3 test set.

ID	nep	ori	pan	spa	swa	tel	tur	urd	zho
M1	0.4592	0.0709	**0.4668**	0.4658	**0.5388**	0.1522	0.4547	0.7483	0.4879
M2	**0.5139**	0.0435	0.16	0.2385	0.3949	0.127	0.3598	**0.7856**	0.511
M3	0.2126	0.0778	0.0688	0.2637	0.1347	0.0953	0.2784	0.1899	0.6915
M4	0.5122	0.0701	**0.4668**	0.4307	0.515	0.1586	0.4633	0.7562	**0.7004**
Baseline	0.1314	**0.3841**	0.4561	**0.5088**	0.2205	**0.6738**	**0.7693**	0.5316	0

Table 7: Macro F1 scores of the four core models on the Subtask 3 test set.

the best. However, in terms of performance on zho (Chinese), M3's performance is the most outstanding. This indicates that although the LLM was fine-tuned using another language, it exhibits substantial cross-lingual transferability.

M4 is the result of the weighted ensembling strategy of this study and also represents our official submission results. While M4 does not necessarily achieve the highest scores across all languages, it demonstrates exceptional proficiency in Chinese. In Subtask 1, this study ranked first in the Chinese language track, indicating that applying an ensembling strategy to multiple well-performing models can lead to further performance enhancements.

5 Conclusion

Through the experimental framework of SemEval-2026 Task 9, this study developed a highly effective system for detecting complex polarized language. Our approach was empirically driven: based on extensive evaluations on the Development data (Configurations D1–D9), we identified that specific LoRA fine-tuning strategies on Llama-3.1 and Qwen2 yielded superior contextual understanding (Configurations D3, D4 and D9). Consequently, we deliberately selected three distinct, high-performing configurations (M1, M2, and M3)—which differ in base architecture, language data, and task integration—to construct a performance-weighted ensemble (M4).

The formal Test set results concretely validate this strategy, demonstrating exceptional predictive power particularly in the Chinese context. Our ensemble model (M4) significantly outperformed the official baseline and secured the 1st place ranking in Subtask 1 (Polarization Detection) for the Chinese (zho). These empirical results prove that strategically ensembling models with diverse inductive biases is highly effective for capturing nuanced sociolinguistic phenomena across varying cultural contexts.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 114-2221-E-324-006.

References

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and

fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

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. *arXiv preprint arXiv:2505.20624*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 487–503.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141.

Teun A Van Dijk. 1998. *Ideology: A multidisciplinary approach*. SAGE Publications Ltd.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682.

Ruth Wodak. 2015. *The politics of fear: What right-wing populist discourses mean*. Sage.

A Development Set Experimental Design and Results

A.1 Experimental Design

Prior to the official model submission, this study conducted comprehensive testing on the Development Set to evaluate the efficacy of models with different architectures in the multilingual polarization detection task. The experiments encompassed the performance of multilingual pre-trained models (e.g., XLM-RoBERTa) and Large Language Models (LLMs) (e.g., Llama-3.1, Qwen2) under various data fine-tuning strategies. Table 8 summarizes the nine sets of experimental configurations (designated D1 to D9) designed in the development set for this study, aiming to comprehensively evaluate the performance of different model architectures, language strategies, and task construction methods when processing multi-contextual tasks. The overall experimental design breaks away from the limitations of a single baseline model, focusing on comparing the "efficacy differences between cross-lingual and monolingual strategies," as well as "performance comparisons among different Large Language Model (LLM) architectures." The specific dimensions of the experimental design are as follows:

A.2 Model Architectures and Training Methods

This study adopted two distinct model paradigms. D1 and D2 employ the encoder-based XLM-RoBERTa for traditional full fine-tuning, serving as one methodological pipeline. D3 through D9 introduce current mainstream decoder-only Large Language Models (LLMs), including different iterative versions of the Qwen series (Qwen2-7B-Instruct, Qwen2.5-7B, Qwen3-8B) and Llama-3.1-8B. To efficiently adapt to downstream tasks under constrained computational budgets, all LLMs utilize LoRA (Low-Rank Adaptation) for Parameter-Efficient Fine-Tuning (PEFT).

A.3 Language Strategies

To verify the effectiveness of different language strategies, we designed three language configurations for the training data:

A.3.1 Cross-Lingual Strategy (all languages)

Configurations such as D2, D4, D5, and D9 mix the training data of all languages for joint training, aiming to explore the models' cross-lingual knowledge

ID	Model	Training data Language	Training data Subtask	Training Method
D1	XLM-RoBERTa	eng	separate 1, 2 & 3	fine-tuning
D2	XLM-RoBERTa	all language	separate 1, 2 & 3	fine-tuning
D3	Qwen2-7B-Instruct	zho	combined 1, 2 & 3	LoRA fine-tuning
D4	Qwen2-7B-Instruct	all language	separate 1, 2 & 3	LoRA fine-tuning
D5	Qwen2.5-7B	all language	separate 1, 2 & 3	LoRA fine-tuning
D6	Qwen3-8B	zho	combined 1, 2 & 3	LoRA fine-tuning
D7	Qwen3-8B	each language	combined 1, 2 & 3	LoRA fine-tuning
D8	Llama3.1-8B	eng	separate 1, 2 & 3	LoRA fine-tuning
D9	Llama3.1-8B	all language	separate 1, 2 & 3	LoRA fine-tuning

Table 8: Employed models, training data configurations, and training methods.

transfer and generalization capabilities.

A.3.2 Specific Monolingual Strategy (eng, zho)

Configurations such as D1 and D8 use exclusively English (eng) data; D3 and D6 use exclusively Chinese (zho) data for fine-tuning.

A.3.3 Comprehensive Monolingual Strategy (each language)

D7 adopts a strategy of independently training a dedicated monolingual model for "each language" in the dataset, thereby conducting a rigorous efficacy comparison with the cross-lingual strategy.

A.4 Subtask Strategies

When addressing the three different subtasks, this experiment compared two construction methods

A.4.1 Task-Decoupled Strategy (Separate 1, 2 & 3)

Most configurations (e.g., D1, D2, D4, D5, D8, D9) adopt the approach of training independent models separately for the three subtasks.

A.4.2 Task-Integrated Strategy (Combined 1, 2 & 3)

D3, D6, and D7 explore a concept similar to multi-task learning, completely merging the training data of the three subtasks to train a unified model to handle the three subtasks simultaneously.

Through this multi-dimensional experimental matrix of D1 to D9, we can deeply analyze which language strategy (cross-lingual joint training vs. independent languages) achieves the optimal balance in this task, and objectively compare the pros and cons of the Qwen family and the Llama-3.1 architectures under different settings.

A.5 Experimental Results

Tables 9 to 14 present the Macro F1 scores for all the aforementioned models across their corresponding languages. Given that our native language is Chinese, this study selected three models (D3, D4, D9) that demonstrate relatively high Macro F1 scores in Chinese and also perform well across other languages to serve as the core models used in our test set. As seen in the tables, D3's performance in Chinese is the most outstanding, while D4 and D9 are the two models with relatively high average scores. Therefore, to compare the differences between identical models with different fine-tuning strategies, as well as different models with identical fine-tuning strategies, this study utilized these three core models in the test set.

B Prompt Templates

In this appendix, we detail the specific prompt templates utilized for each of the core models (M1, M2, and M3). Recognizing that models with different base architectures (Llama-3.1 vs. Qwen2) and fine-tuning paradigms (Task-Decoupled vs. Task-Integrated) exhibit varying instruction-following behaviors, we tailored the prompt formulations to optimize the feature extraction capabilities of each specific model configuration. The input texts are denoted by the placeholder `{Input Text}`.

B.1 Prompts for M1: Llama-3.1-8B-Instruct (Task-Decoupled)

For M1, which employs a cross-lingual task-specific fine-tuning strategy, the following independent prompts were utilized for the three subtasks:

Subtask 1: Polarization Detection

System Prompt: [
"你是一個文字分類器。"

ID	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
D1	0.72	0.6816	0.8315	0.691	0.7229	0.8274	0.4709	0.725	0.3688	**0.61**	**0.8881**
D2	**0.7425**	0.7943	0.8154	0.6916	0.7724	**0.8724**	0.7729	0.7658	0.6602	0.6305	0.8725
D3	0.2913	0.7495	0.7672	0.6914	0.7587	0.4557	0.5081	0.6211	0.4864	0.0995	0.318
D4	0.5019	**0.8178**	0.8181	0.7483	**0.8077**	0.84	**0.7945**	0.7365	0.6458	0.5057	0.8218
D5	0.587	0.8153	0.8655	**0.767**	0.7986	0.8274	0.7501	**0.791**	0.5885	0.5534	0.7571
D6	0.4757	0.8145	0.8	0.7346	0.6437	0.5658	0.5246	0.7697	0.4169	0.1498	0.6972
D7	0.7099	0.7962	**0.8962**	0.7546	0.7743	0.8333	0.7943	0.775	0.6352	0.5897	0.8708
D8	0.3378	0.5834	0.5802	0.5865	0.7862	0.4287	0.5036	0.5627	0.5768	0.2312	0.386
D9	0.4831	0.52	0.4926	0.5345	0.696	0.6822	0.4836	0.3851	**0.6843**	0.4547	0.7005

Table 9: Macro F1 scores across various languages on the Subtask 1 development set.

ID	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
D1	0.87	0.7274	0.6862	0.7538	0.683	0.6662	0.7737	0.8641	0.7477	0.7186	0.8692
D2	**0.88**	0.7728	0.8298	0.7789	0.7236	0.6951	0.7765	0.8214	0.7561	0.7383	0.8551
D3	0.5783	0.5616	0.4792	0.6973	0.6401	0.6886	0.4783	0.4373	0.7207	0.557	**0.9439**
D4	0.8598	0.5623	0.7494	0.781	0.76	0.6964	0.8106	0.8301	0.7651	0.6713	0.9299
D5	0.8098	0.6878	0.719	0.8448	**0.8185**	0.7439	0.8108	0.8641	0.7999	0.7521	0.9159
D6	0.7672	0.7374	0.6656	0.7438	0.6166	0.67	0.5745	0.6767	**0.8346**	0.6467	0.9299
D7	0.85	**0.8738**	0.8399	**0.8538**	0.8184	0.7182	0.8161	**0.9067**	0.8087	**0.7659**	0.9159
D8	0.4831	0.7986	**0.8714**	0.7667	0.794	**0.8333**	0.6574	0.7655	0.6198	0.5325	0.8307
D9	0.86	0.7037	0.7689	0.8521	0.7988	0.7325	**0.8216**	0.8808	0.8258	**0.7659**	0.9112

Table 10: Macro F1 scores across various languages on the Subtask 1 development set.

"請根據輸入的 text 判斷 polarization。"
"你只能輸出一個字元：0 或 1。"
"不要輸出任何解釋、不要輸出其他字。"
]
(English Translation: [
"You are a text classifier."
"Please determine the polarization based on the input text."
"You must only output a single character: 0 or 1."
"Do not output any explanations, do not output any other words."
])
User Prompt: [{Input Text}]

Subtask 2: Polarization Type Classification

System Prompt: [
"你是一個多標籤文字分類器（multi-label）。"
"請根據輸入的 text 判斷以下 5 個類別是否出現（0 或 1）："
"political, racial/ethnic, religious, gender/sexual, other。"
"你只能輸出一行，格式必須完全如下："
"political=<0/1>, racial/ethnic=<0/1>, religious=<0/1>, gender/sexual=<0/1>, other=<0/1>"
"不要輸出任何解釋、不要輸出其他字。"
]
(English Translation: [
"You are a multi-label text classifier."
"Please determine whether the following 5 categories appear (0 or 1) based on the input text:"
"political, racial/ethnic, religious, gender/sexual, other."
"You must only output a single line, and the format must be exactly as follows:"
"political=<0/1>, racial/ethnic=<0/1>, religious=<0/1>, gender/sexual=<0/1>, other=<0/1>"
"Do not output any explanations, do not output any other words."
])
User Prompt: [{Input Text}]

Subtask 3: Manifestations Classification

System Prompt: ["你是一個多標籤文字分類器。"
"請根據輸入的 text，判斷以下 6 種仇恨或極化表現是否存在（0 或 1）："
"stereotype, vilification, dehumanization, extreme_language, "
"lack_of_empathy, invalidation。"
"你只能輸出一行，格式必須完全如下："

ID	amh	arb	ben	deu	eng	fas	hau	hin	ita	khm	mya
D1	0.4551	0.417	0.1994	0.3756	0.2739	**0.5783**	0.0727	0.7362	**0.4163**	0.6883	0.4555
D2	0.4179	0.4405	0.1719	0.3424	0.38	0.5446	0.1444	**0.7955**	0.392	**0.7074**	**0.5073**
D3	0.0336	0.4821	**0.3137**	0.4588	**0.4652**	0.2489	0.0805	0.341	0.234	0.1207	0.0061
D4	0.3109	0.5514	0.2866	0.4541	0.3914	0.5455	0.125	0.7787	0.3949	0.358	0.4237
D5	0.4015	**0.6605**	0.2477	0.4954	0.3801	0.4137	0.1111	0.7823	0.3955	0.4922	0.4773
D6	0.2196	0.4983	0.2933	0.4756	0.3682	0.324	0.0587	0.4266	0.2388	0.2221	0.1988
D7	**0.4566**	0.604	0.2717	**0.5398**	0.382	0.5341	0.4434	0.7544	0.369	0.6841	0.487
D8	0.1809	0.0837	0.125	0.0372	0.15	0.069	0	0.1831	0.0933	0.1548	0.1525
D9	0.3027	0.5062	0.1463	0.4144	0.3906	0.4184	**0.8221**	0.5118	0.2737	0.226	0.2687

Table 11: Macro F1 scores across various languages on the Subtask 2 development set.

ID	nep	ori	pan	pol	rus	spa	swa	tel	tur	urd	zho
D1	**0.8056**	0.2083	0.2878	0.4057	0.5093	0.5792	0.4198	0	0.4542	0.6526	0.7245
D2	0.6902	0.4701	0.2878	0.4733	0.4413	0.5528	**0.5275**	0.3138	0.4791	0.6902	0.6396
D3	0.415	0.2128	0.0857	0.4646	0.4939	0.5229	0.153	0.0733	0.4452	0.2262	0.8273
D4	0.7401	0.4995	0.2656	0.582	0.5417	0.5668	0.4533	0.1254	0.5873	0.7495	0.7671
D5	0.746	0.3133	0.2483	0.5465	0.5687	0.5951	0.4349	0.1317	0.5215	0.7357	**0.8614**
D6	0.7029	0.4853	0.3196	0.5352	0.4733	0.5303	0.2353	0.2314	0.4829	0.3193	0.74
D7	0.7498	**0.7374**	0.3954	**0.6204**	**0.6187**	**0.6179**	0.4489	**0.3342**	**0.612**	**0.7908**	0.7763
D8	0.0571	0.1829	0	0.0255	0.0943	0.1379	0.0031	0.1505	0.092	0.0768	0.0873
D9	0.4603	0.1429	**0.4006**	0.429	0.3813	0.5268	0.1733	0.2824	0.4506	0.7253	0.5117

Table 12: Macro F1 scores across various languages on the Subtask 2 development set.

"stereotype=<0/1>, vilification=<0/1>, dehumanization=<0/1>, "
"extreme_language=<0/1>, lack_of_empathy=<0/1>, invalidation=<0/1>"
"不要輸出任何解釋、不要輸出其他字。"]
(English Translation: [
"You are a multi-label text classifier."
"Please determine whether the following 6 types of hateful or polarized manifestations exist (0 or 1) based on the input text:"
"stereotype, vilification, dehumanization, extreme_language, "
"lack_of_empathy, invalidation."
"You must only output a single line, and the format must be exactly as follows:"
"stereotype=<0/1>, vilification=<0/1>, dehumanization=<0/1>, "
"extreme_language=<0/1>, lack_of_empathy=<0/1>, invalidation=<0/1>"
"Do not output any explanations, do not output any other words."
])
User Prompt: [{Input Text}]

B.2 Prompts for M2: Qwen2-7B-Instruct (Task-Decoupled)

Similarly utilizing a cross-lingual task-specific strategy, M2 (based on the Qwen2 architecture) was guided by the following distinct set of prompts:

Subtask 1: Polarization Detection

System Prompt: ["Task 1: Binary Polarization Classification (0 = Not Polarized, 1 = Polarized). A text is polarized if it expresses divisive, hostile, or extreme attitudes toward out-groups, or shows blind support toward in-groups. Polarization includes stereotyping, vilification, dehumanization, extreme or absolutist language, intolerance of others' identities or viewpoints, or calls for harm. The input text may appear in any language; apply the same criteria regardless of language. Do not rely on language-specific cues—base the decision only on the meaning and intent of the text. If any polarized characteristics appear, output 1; otherwise output 0. Always judge the overall meaning and context, not isolated words. Return only a single label: 0 or 1."]
User Prompt: [{Input Text}]

Subtask 2: Polarization Type Classification

System Prompt: ["Task 2: Polarization Type Classification (Multi-label Output).

ID	amh	arb	ben	deu	eng	fas	hau	hin	khm
D1	0.3145	0.3118	0.1	0.1852	0.189	0.2454	0	0.7323	0.2165
D2	0.3309	0.4457	0.0905	0.2883	0.3979	0.2493	0	0.7327	**0.2961**
D3	0.0408	0.3091	0.1925	0.2826	0.3197	0.1558	0.0202	0.2645	0.0159
D4	0.2308	0.4979	0.0997	0.365	0.3386	0.2411	0	**0.7633**	0.1356
D5	0.2693	0.4837	0.087	0.424	0.3679	0.2551	0	0.7376	0.1601
D6	0.1268	0.2969	**0.2493**	0.2952	0.2923	0.1671	0.0833	0.2873	0.0651
D7	0.3126	**0.5631**	0.112	**0.4858**	0.4307	**0.146**	0.146	0.7622	0.2844
D8	0.3007	0.4754	0.138	0.4144	**0.4414**	**0.2886**	0.047	0.6502	0.1236
D9	**0.3678**	0.5287	0.0414	0.4015	0.3489	0.2385	0	0.71	0.1528

Table 13: Macro F1 scores across various languages on the Subtask 3 development set.

ID	nep	ori	pan	spa	swa	tel	tur	urd	zho
D1	0.5759	0	0.3356	0.3368	**0.5088**	0.1061	0.319	0.7442	0.5422
D2	0.6271	0.119	0.3356	0.3779	0.3874	0.2704	0.3844	0.732	0.5319
D3	0.1038	**0.2226**	0.0762	0.2291	0.0841	0.0751	0.2494	0.1743	**0.7243**
D4	0.5717	0.0222	0.3131	0.4162	0.4679	0.1864	0.4285	0.8085	0.6184
D5	0.6229	0.0996	0.4051	0.4466	0.4554	0.2118	**0.458**	0.7263	0.6861
D6	0.3024	0.0714	0.1726	0.2362	0.1262	0.163	0.1958	0.2297	0.6073
D7	**0.6428**	0.1763	**0.5298**	**0.4592**	0.4624	**0.3085**	0.4059	**0.8122**	0.714
D8	0.3644	0.0831	0.45	0.3631	0.411	0.1459	0.4164	0.7169	0.333
D9	0.4687	0.0469	0.4145	0.3617	0.43	0.2296	0.4503	0.7234	0.5519

Table 14: Macro F1 scores across various languages on the Subtask 3 development set.

Given a social media text, identify which types of polarization it targets. A text may contain zero, one, or multiple types. Return a 5-label binary vector in the following fixed order:
1. Political/ideological polarization
2. Racial or ethnic polarization
3. Religious polarization
4. Gender or sexual identity polarization
5. Other types of polarization (economic, media, technology, etc.)
Output format: five comma-separated digits (each 0 or 1), e.g., "0,1,0,0,0".
Guidelines:
- Select a type if the text contains direct or implied hostility, intolerance, stereotyping, blame, or conflict targeting that group.
- Multiple labels may apply simultaneously.
- If none apply, output "0,0,0,0,0".
- The input text may be in any language; apply the same criteria regardless of language.
- Return only the 5 numeric labels, nothing else."]

User Prompt: [{Input Text}]

Subtask 3: Manifestations Classification

System Prompt: ["Task 3: Polarization Manifestation Classification (Multi-label Output).
Given a social media text, identify which rhetorical tactics or manifestations of polarization appear in the message. A text may contain zero, one, or multiple manifestations. Return a 6-label binary vector in the following fixed order:
1. Stereotype
2. Vilification
3. Dehumanization
4. Extreme language and absolutism
5. Lack of empathy or understanding
6. Invalidation
Output format: six comma-separated digits (each 0 or 1), e.g., "1,0,0,1,0,0".
Guidelines:
- Select a label if the text contains the corresponding rhetorical tactic.
- Multiple labels may apply simultaneously.
- If none apply, output "0,0,0,0,0,0".

- The input text may be in any language; apply the same criteria consistently.
- Return only the six numeric labels, nothing else."]
User Prompt: [{Input Text}]

B.3 Prompt for M3: Qwen2-7B-Instruct (Task-Integrated)

For M3, which adopts a monolingual (Chinese) multi-task joint fine-tuning strategy, we designed a unified prompt to concurrently extract labels across all three subtasks.

Unified Multi-Task Prompt (Chinese)

System Prompt: [你是一位精通多國語言的社會語言學專家。
目前正在分析的語言是：中文(Chinese)。
請根據該語言的文化背景，判斷以下句子是否包含極化語言及特定類別。
請直接輸出 JSON 格式結果。]
*(English Translation: [You are a sociolinguistic expert proficient in multiple languages.
The language currently being analyzed is: Chinese.
Please determine whether the following sentence contains polarized language and specific categories based on the cultural background of this language.
Please output the results directly in JSON format.])*

User Prompt: [你是一位精通多國語言的社會語言學專家。
目前正在分析的語言是：**[**動態填入語言**]**。請根據該語言的文化背景，判斷以下句子是否包含極化語言及特定類別。
請直接以 JSON 格式輸出結果，不要包含任何多餘的解釋。 {Input Text}]
*(English Translation: [You are a sociolinguistic expert proficient in multiple languages.
The language currently being analyzed is: **[Dynamically inserted language]**.
Please determine whether the following sentence contains polarized language and specific categories based on the cultural background of this language.
Please output the results directly in JSON format, do not include any redundant explanations. {Input Text}])*