Team QUST at SemEval-2025 Task 10: Evaluating Large Language Models in Multiclass Multi-label Classification of News Entity Framing

Jiyan Liu[1], Youzheng Liu[1], Taihang Wang[1], Xiaoman Xu[1]*, Yimin Wang[2] and Ye Jiang[1]
College of Information Science and Technology[1]
College of Data Science[2]
Qingdao University of Science and Technology, China

Abstract

This paper introduces the participation of the QUST team in subtask 1 of SemEval-2025 Task 10. We evaluate various large language models (LLMs) based on instruction tuning (IT) on subtask 1. Specifically, we first analyze the data statistics, suggesting that the imbalance of label distribution made it difficult for LLMs to be fine-tuned. Subsequently, a voting mechanism is utilized on the predictions of the top-3 models to derive the final submission results. The team participated in all language tracks, achieving 1st place in Hindi (HI), 2nd in Russian (RU), 3rd in Portuguese (PT), 6th in Bulgarian (BG), and 7th in English (EN) on the official test set. We release our system code at: https://github.com/warmth27/SemEval2025_Task10

1 Introduction

SemEval-2025 Task 10 encourages participants to develop algorithms for multilingual recognition and extraction of narrative tasks from online news (Piskorski et al., 2025; Stefanovitch et al., 2025). We participated in subtask 1 (Entity Framing), which aims to assign fine-grained role labels, covering three main types (protagonists, antagonists, and innocent) to named entities (NEs) mentioned in news articles, based on a predefined fine-grained role classification system (Mahmoud et al., 2025). This is a multi-label multi-class text-span classification task.

During this process, we face several challenges:

- **Complexity of languages:** Compared to widely spoken languages like English, smaller languages such as Bulgarian lack high-quality pre-trained language models, which significantly heighten the complexity of model architecture design.
- **Data scarcity:** Insufficient training data has resulted in suboptimal fine-tuning outcomes for the large language models (LLMs). Performance is also weaker for languages with smaller datasets.

In subtask 1, we first conduct a quick evaluation by comparing our model with the baseline model. Inspired by Xu et al. (2024) and Zhang et al. (2023), we apply instruction tuning (IT) to several LLMs on subtask 1. However, IT entails substantial training costs. We first fine-tuning the models in English, then apply it to other languages to minimize fine-tuning and evaluation time.

To enhance the model's performance and generalization ability, we adopt a hard voting based on ensemble learning (Jabbar, 2024), in which the top-3 selected models vote to determine the final prediction. Experimental results indicate that ensemble learning significantly enhances the model performance.

2 Related Work

2.1 Instruction tuning for LLMs

IT is a powerful technique that adjusts the input context to align with specific instructions, updating the parameters of LLMs in a supervised manner (Wang et al., 2024b; Jiang, 2023). Current studies frequently examine the efficacy of IT for LLMs. For instance, Qin et al. (2024) presents a comprehensive review of IT in LLMs, detailing the fine-tuning process with instruction pairs and evaluating the critical factors that influence the outcomes of IT. Wang et al. (2024a) emphasizes that in the IT process of LLMs, the quality of the dataset plays a more significant role than its quantity.

Similar to the aforementioned studies, in subtask 1, IT helps the model understand complex contextual information and accurately assign fine-grained roles to entities based on instructions. By applying customized IT, we can improve the model's

*Corresponding author

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1052–1056

performance on complex tasks, particularly when large-scale labeled data is scarce, as this method effectively enhances the model's generalization ability.

2.2 Voting

Voting is an ensemble learning strategy that enhances overall performance by combining the predictions of several base models (Xu et al., 2024; Abro, 2021). The voting strategy we adopt is hard voting, in which the final result is determined by the majority vote based on the predicted class labels of the top-3 models. The most frequent class is chosen as the final outcome.

Language	train	dev	test
BG	627	31	124
EN	686	91	235
HI	2331	280	316
PT	1251	116	297
RU	722	86	214

Table 1: Statistics of each language

3 Experimental Setup

3.1 Data

The statistical distribution of each language dataset is presented in Table 1. It is evident that the data distribution is imbalanced across the five languages, especially for languages with fewer training samples (such as BG and EN), which could impact the model's generalization ability.

Additionally, we analyze the distribution of label counts, as shown in Figure 1. The results indicate that the majority of samples are single-labeled. Based on this observation, we design an experiment in which the task is treated as a single-label task to evaluate whether this approach can enhance the original results. The experimental results and analysis are presented in Section 4.1.

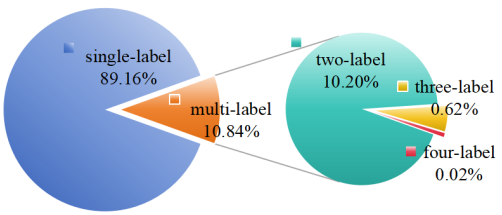

Figure 1: Distribution of the number of labels in the dataset.

3.2 Instruction strategy

Inspired by Wang et al. (2024b), we improve several versions based on their work. Ultimately, we find that this version of the instruction yields the best results. The example directive is: "*Given an article and an entity within that article. Analyze this article and the entity, and provide the fine-grained roles of the entity.*" This instruction emphasizes the requirement to assignone or more fine-grained roles to each entity, supporting a one-to-many label output.

3.3 Model configurations

Before identifying the final models, We first compared the performance of several baseline models from Hugging Face[1]. Considering the multilingual nature of the task, we primarily conduct experiments using the English dataset during the model comparison phase. The models involved in the evaluation include DeBERTa-v3-small (He et al., 2021), GLM4-9B-chat (GLM et al., 2024), Qwen2-7B-instruct (Yang et al., 2024), Qwen2.5-14B-instruct (Yang et al., 2024), Phi-3-small-128k-instruct (Phi-3-small) (Abdin et al., 2024a), Phi-3-medium-128k-instruct (Phi-3-medium) (Abdin et al., 2024a), and Phi-4 (Abdin et al., 2024b).

Given the variations in data size across languages, as well as considerations for training time and computational efficiency, we established two training schemes: 10 epochs and 20 epochs. The learning rate for the Qwen2-7B-instruct model is set to 1e-5, while for the other models, it is set to 1e-4. To avoid overfitting and enhance storage efficiency, we implement an epoch selection strategy, retaining only the model parameters that yield the best performance.

The final result utilizes an ensemble learning strategy, employing hard voting to combine the predictions of the top-3 selected models for each language, thereby enhancing the prediction accuracy of the final result.

4 Results

4.1 Single-label result

Since the majority of samples are single-labeled, we extract data containing only single labels from

[1]https://huggingface.co/

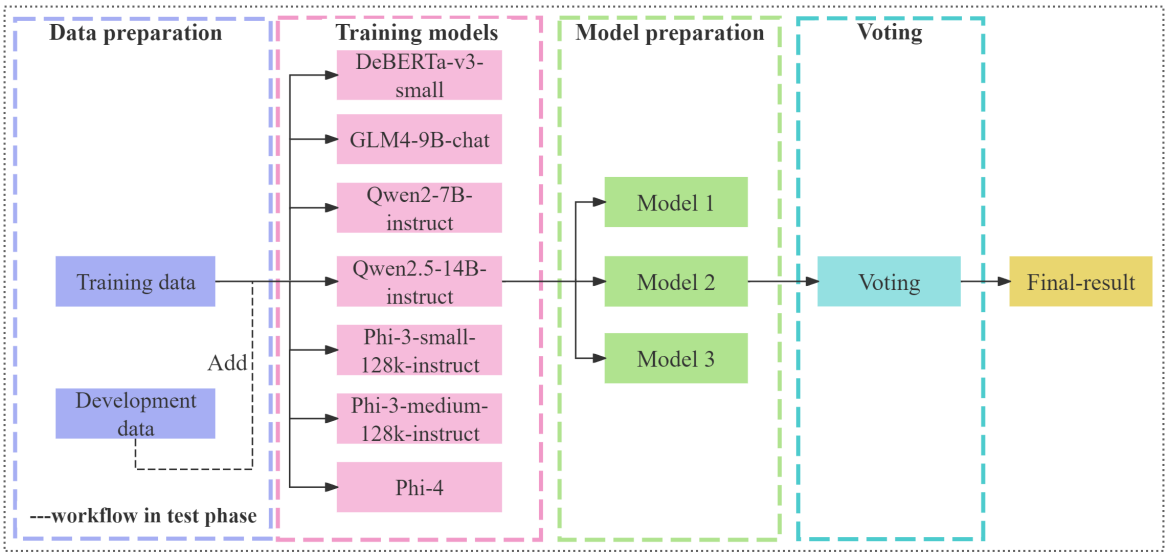

Figure 2: llustration of the overall workflow in this paper. "Model 1", "Model 2" and "Model 3" represent the top-3 best models of the performance employed.

Language	Model	Methods	Exact Match Ratio
EN	Phi-3-small-128k-instruct	Single-label	0.3736
		Multi-label	**0.3846**

Table 2: The result of Single-label Vs. Multi-label. **Bold** indicates the best result.

the English training dataset to treat this task as a single label task. As shown in Table 2, the score for the single-label task is 0.3736, whereas the score for the multi-label task is 0.3846. The result for the single-label task drops by only 2.86%. Although this approach does not improve model performance, the decline is minimal, suggesting a significant imbalance in the dataset.

4.2 Evaluation results

Our evaluation primarily relies on the official development data. The goal of our evaluation is to rapidly identify models and methods that perform well, and to compare the performance variations among different models. To enhance experimental efficiency, we focus primarily on evaluating the English dataset at this stage.

As shown in Table 3, all large language models except GLM4-9B-chat significantly outperform DeBERTa-v3-small, confirming that IT can effectively enhance the performance of LLMs on this task. Among these, GLM4-9B-chat yields the lowest result, possibly due to limited instruction comprehension, which leads to weaker generalization ability.

Phi-4 achieves the best performance among the single models, slightly outperforming Phi-3-medium and Phi-3-small, suggesting that larger-scale Phi-series models offer more stability on this task.

Voting strategy further enhances the final performance by 1.1% compared to the Phi-4, indicating that the voting strategy effectively integrates the advantages of the individual models. However, the improvement is limited, possibly due to similar predictions across models or the inherent performance ceiling of the task.

4.3 Official test results

Our approach participated in several language tracks, with Hindi ranked 1st, Russian ranked 2nd, Portuguese ranked 3rd, and Bulgarian and English ranked 6th and 7th, respectively, as shown in Table 4. Additionally, our approach significantly outperforms the baseline of subtask 1 across all languages, demonstrating the effectiveness of our approach. However, it is worth noting that our performance in English and Bulgarian is comparatively weaker, while the final test results for Hindi are better than the performance in the previous evaluation phase.

This unexpected phenomenon may be related to the scale of the training data. The training data for

Methods	Models	Exact Match Ratio
Small language model	DeBERTa-v3-small	0.2747
Instruction tuning	GLM4-9B-chat	0.1978
	Qwen2-7B-instruct	0.3626
	Qwen2.5-14B-instruct	0.3626
	Phi-3-small-128k-instruct	0.4505
	Phi-3-medium-128k-instruct	0.4505
	Phi-4	<u>0.4615</u>
Voting	Phi-3-small+Phi-3-medium+Phi4	**0.4725**

Table 3: Evaluation of methods and models on English development data. **Bold** indicates the best result, and <u>Underline</u> indicates the second-best result.

Language	Baseline subtask 1	QUST(rank) subtask 1
BG	0. 0403	0. 3871(6th)
EN	0. 0383	0. 3277(7th)
HI	0. 0570	**0. 4684(1st)**
PT	0. 0471	0. 4579(3rd)
RU	0. 0514	<u>0. 5140(2nd)</u>

Table 4: Official test results.**Bold** indicates the best result, and <u>Underline</u> indicates the second-best result.

English and Bulgarian is relatively limited, which may limit the model's generalization ability, while Hindi benefits from a richer dataset, allowing the model to better learn task patterns and perform more effectively in the test phase. Furthermore, Table 4 shows strong performance in Russian and Portuguese, further indicating that the scale of training data might be a key factor affecting model performance.

5 Conclusion

In summary, our team developed an effective approach for subtask 1 of SemEval-2025 Task 10. We first conduct instruction tuning on large language models on the English dataset and choose the models that perform best, then adapt them to other languages. In the final testing phase, to augment the training data, we incorporate the development data into the training set and select the top-performing model based on evaluation. Ultimately, hard voting successfully integrates the advantages of several models, thereby enhancing the prediction accuracy.

As the training data for English and Bulgarian is relatively limited and we have not yet employed data augmentation methods, future work will explore effective strategies to augment both the quantity and quality of data, thereby enhancing the model's capacity to comprehend texts from diverse languages and cultural backgrounds.

Acknowledgements

This work is funded by the Natural Science Foundation of Shandong Province under grant ZR2023QF151 and the Natural Science Foundation of China under grant 12303103.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Abdul Ahad Abro. 2021. Vote-based: Ensemble approach. *Sakarya University Journal of Science*, 25(3):858–866.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Hanan Ghali Jabbar. 2024. Advanced threat detection using soft and hard voting techniques in ensemble learning. *Journal of Robotics and Control (JRC)*, 5(4):1104–1116.

Ye Jiang. 2023. Team qust at semeval-2023 task 3: A comprehensive study of monolingual and multilingual approaches for detecting online news genre, framing and persuasion techniques. *arXiv preprint arXiv:2304.04190*.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, et al. 2025. Entity framing and role portrayal in the news. *arXiv preprint arXiv:2502.14718*.

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

Yulei Qin, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li, and Xing Sun. 2024. Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. *arXiv preprint arXiv:2408.02085*.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024a. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*.

Taihang Wang, Xiaoman Xu, Yimin Wang, and Ye Jiang. 2024b. Instruction tuning vs. in-context learning: revisiting large language models in few-shot computational social science. *arXiv preprint arXiv:2409.14673*.

Xiaoman Xu, Xiangrun Li, Taihang Wang, Jianxiang Tian, and Ye Jiang. 2024. Team qust at semeval-2024 task 8: A comprehensive study of monolingual and multilingual approaches for detecting ai-generated text. *arXiv preprint arXiv:2402.11934*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.