YNU-HPCC at SemEval-2026 Task 6: Hierarchical Taxonomy Prompting and CoT Distillation for Political Clarity Classification

Canning Wen, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: wencn@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

In political interviews, politicians frequently employ evasion strategies to avoid direct answers, making it challenging to evaluate response clarity in Natural Language Processing. This paper presents the YNU-HPCC system for SemEval-2026 task 6: clarity classification in political interviews. To address the limitation where traditional models capture only surface-level semantics, this paper proposes two reasoning-enhanced frameworks. First, we introduce Hierarchical Taxonomy Prompting. This method guides LLMs to follow a strict top-down classification logic. Specifically, the model determines the clarity level before identifying specific evasion techniques. Furthermore, it explicitly articulates the reasoning process. Second, to balance reasoning capability with resource constraints, we employ Chain-of-Thought Distillation. We utilize DeepSeek V3.1 as a teacher model to generate comprehensive reasoning chains, which are then used to SFT the smaller student models. Experimental results demonstrate the effectiveness of our approach: The system achieved 6th place in Task 1 and 5th place in Task 2 among all participating teams, highlighting the importance of reasoning processes in detecting complex linguistic evasion. The code can be found at: https://github.com/wencanning/SemEval2026-Task6

1 Introduction

In recent political interviews, politicians increasingly tend to give ambiguous responses and deflect from topics (Thomas et al., 2024). To discern the true intent behind such rhetoric, employing natural language processing (NLP) techniques (Shen et al., 2026) for implication detection has become imperative. However, prior research has scarcely addressed the evaluation of political clarity. This task bridges the gap.

Task 6 in the SemEval-2026 (Thomas et al., 2026) is a classification task. It consists of two subtasks.

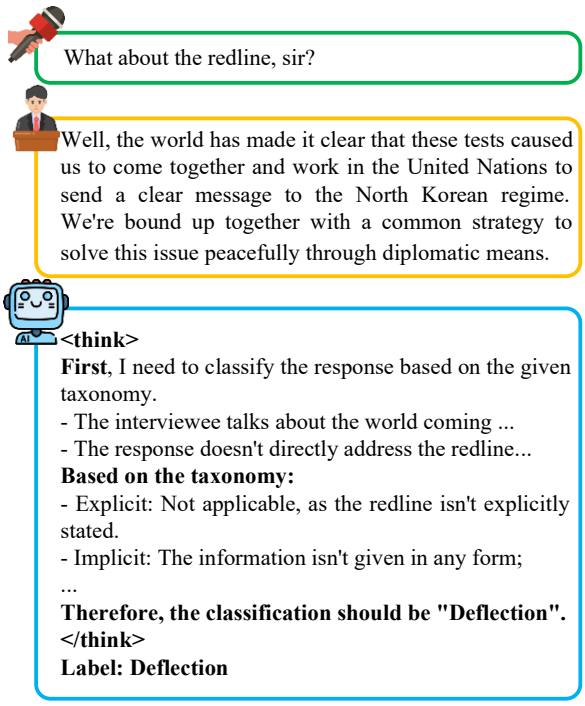

Figure 1: An example of how LLM takes the evaluation of response clarity.

- Task 1: This task has 3 labels. We need to identify the clarity level of the interview question answer pair, which is Clear Reply, Ambivalent, and Clear Non-Reply.
- Task 2: This task has 9 labels. Based on the clarity label, further classify the evasion strategy the politician used.

In previous political clarity classification tasks, researchers used simple prompt engineering and supervised finetuning (SFT) to study clarity levels and evasion strategies. However, these traditional Prompt Engineering methods, such as one-shot, few-shot, and SFT, only instruct the LLM to output the final label. Because the LLM lacks token-level

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 308–313

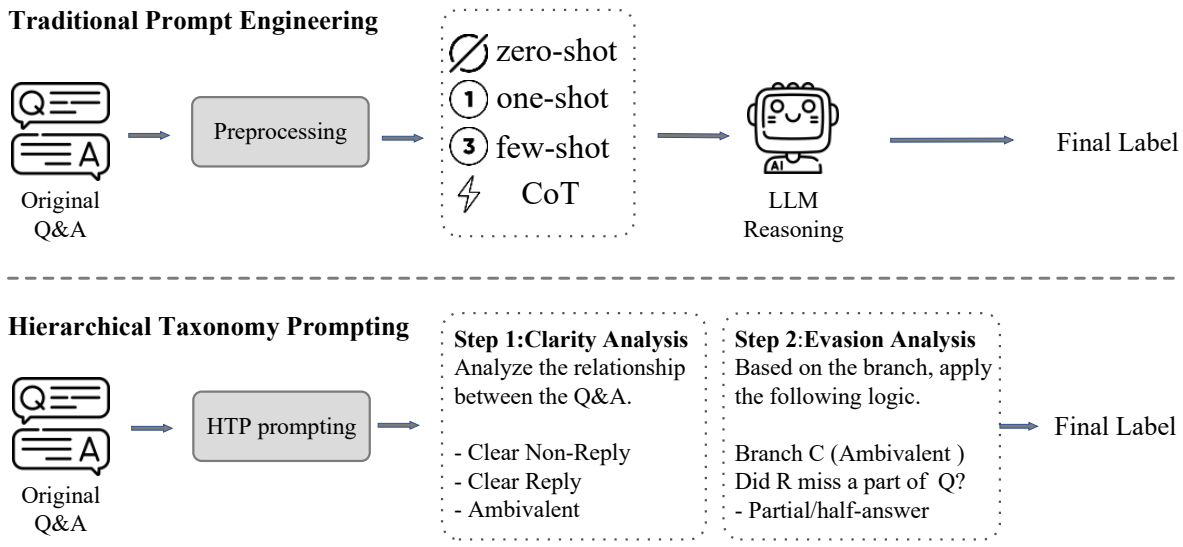

Figure 2: Comparison between traditional prompt engineering and the proposed HTP framework.

consumption, it only understands surface-level semantic information in the text. This led previous work to struggle with the classification accuracy. In contrast, we force the LLM to output the analysis (Guo et al., 2025) before the final label. By producing a long chain-of-thought (CoT) (Zheng et al., 2024), the LLM can analyze the interview question-answer pair more carefully.

The study primarily employs two methodological frameworks to classify question clarity and evasion techniques. The first approach is based on prompting. We adopt hierarchical taxonomy prompting (HTP), instructing the LLM to strictly adhere to the taxonomy proposed by the task organizers. Specifically, the model first classifies the clarity level to determine the high-level category. Based on this high-level classification, the LLM then identifies the specific evasion technique employed by the politician. To enhance classification accuracy, we require the LLM to articulate the reasoning process (Chen et al., 2024) for determining both label levels before generating the final output. The second approach is based on finetuning. To fully leverage the capabilities of reasoning models while adhering to resource constraints, we opt for CoT distillation on smaller models. Specifically, we employ DeepSeek V3.1 (Liu et al., 2024) as the teacher model, while Qwen3 (Yang et al., 2025) and the Llama 3.1 series (Grattafiori et al., 2024) serve as student models. We input the QEvasion training set into DeepSeek using designed prompt templates to generate comprehensive reasoning, extracting these chains as distillation data. Finally, the smaller models (Chen et al., 2025) perform SFT on the extracted reasoning chains.

2 Related Work

CoT Prompting. Chain-of-Thought prompting guides models to explicitly generate intermediate reasoning steps before producing the final output. Extensive prior work has demonstrated that engaging in such comprehensive reasoning significantly enhances LLM performance (Wei et al., 2022). In the system, comparisons between models that incorporate reasoning processes and those that directly predict answers reveal performance gains of up to 9 percentage points.

Knowledge Distillation. Knowledge Distillation (Hinton et al., 2015) is primarily employed in resource-constrained scenarios. The standard paradigm involves selecting a powerful, large-scale model as the teacher and a smaller model as the student with appropriate parameters. The student model aims to mimic the teacher's outputs, thereby acquiring the teacher's capabilities. In our approach, we perform SFT on the student model using high-quality Chain-of-Thought data generated by DeepSeek V3.1, a method often referred to as black-box distillation.

3 Overview of System

3.1 Hierarchical Taxonomy Prompting

We designed a structured prompt to guide the LLM's reasoning process. This prompt strictly

Method	Model	Task 1		Task 2	
		Acc	F1	Acc	F1
HTP	MiMo-V2-Flash	0.814	**0.776**	0.529	**0.559**
Fine Tuning	DeBERTa-base	0.256	0.136	0.373	0.06
	RoBERTa-base	0.405	0.402	0.431	0.258
	ModernBERT-base	0.610	0.534	0.457	0.344
	DeBERTa-large	0.334	0.228	0.428	0.114
	RoBERTa-large	0.256	0.136	0.373	0.060
	ModernBERT-large	0.717	0.633	**0.529**	0.476
SFT Only	Qwen3-14B	0.720	0.668	0.526	0.411
	Llama-3.1-8B	0.646	0.529	0.490	0.285
SFT on DeepSeek CoT	Qwen3-4B	0.789	0.703	0.405	0.415
	Qwen3-8B	0.785	0.672	0.399	0.423
	Qwen3-14B	0.792	0.715	0.445	0.509
	Llama-3.1-8B	**0.824**	0.753	0.448	0.489

Table 1: Results of different methods and models on Task 1 and Task 2. The best performance within each method category is underlined, while the best overall performance across all methods is bolded.

adheres to the hierarchical taxonomy. It forces the model to execute a two-step decision tree workflow.

First, the prompt defines the classification rules. It explicitly instructs the model to determine the clarity label in step 1. The model must analyze whether the response is Clear Reply, Clear Non-Reply, or Ambivalent Reply. This step filters the response into one of three main branches. Second, the prompt guides the model to identify the evasion label in step 2. The logic depends on the branch selected in step 1. For Ambivalent replies, we incorporated specific logical tests into the prompt. These include a logic check for implicit answers, a completeness check for partial answers, and a pivot test to distinguish between deflection and dodging. Finally, to ensure the reasoning process is explicit, the prompt enforces a strict output format. The model must generate a textual macro analysis and micro analysis before outputting the final label. This mechanism ensures that the model performs a detailed analysis before reaching a conclusion.

$$\text{Ana}, \text{Label} = \text{LLM}(\text{P}(\text{Q},\text{SubQ},\text{Ans})) \quad (1)$$

The entire process can be formalized as Equation 1, where Q denotes the journalist's question, SubQ denotes the specific sub-question addressed by the politician (since the initial question may contain multiple parts), and Ans denotes the politician's response. P represents the prompt template. Ana and Label refer to the detailed analysis and the final predicted evasion level label generated by the LLM, respectively.

3.2 Finetuning and CoT Distillation

We conducted finetuning on both Encoder-Only and Decoder-Only models to compare the performance of different architectures on this task.

For Encoder-Only models, we selected a series of BERT-based models (Devlin et al., 2019) for experimentation. We first concatenate the journalist's question, the politician's response, and the specific sub-question addressed. This sequence is then fed into BERT to extract feature vectors. Finally, a fully connected layer maps the extracted feature vector to a 9-dimensional vector, and the final label is obtained via softmax and argmax operations. For Decoder-Only models, we utilized the Qwen3 and Llama 3.1 model series. To enhance the reasoning capabilities of these smaller models, we implemented CoT distillation using DeepSeek. First, we input the interview Q&A pairs into DeepSeek using a carefully designed prompt to collect both the reasoning process and the final evasion label, resulting in an augmented dataset in the format (prompt, reasoning, label). Subsequently, we performed SFT on the smaller models using this dataset, where the reasoning process is encapsulated within <think> </think> XML tags. Post SFT, when presented with a Q&A pair, the model explicitly generates the reasoning process wrapped in tags before outputting the final evasion label.

Method	Task 1		Task 2	
	Acc	F1	Acc	F1
zero-shot	0.799	0.742	0.456	0.486
one-shot	0.799	0.751	0.493	0.490
few-shot	0.799	0.734	0.513	0.503
CoT	0.792	0.733	0.435	0.470
HTP	**0.814**	**0.776**	**0.529**	**0.559**

Table 2: Performance comparison of MiMo-V2-Flash across different prompt engineering methods. The best performing metrics are highlighted in bold.

4 Experiment Details

Dataset. We exclusively utilized the QEvasion dataset[1] provided by the task organizers. The most critical fields utilized in our experiments are interview_question and interview_answer. All data entries are in English. The dataset consists of 3,450 training samples and 308 test samples.

Model Selection. For the prompting-based approach, we employed the MiMo-V2-Flash (Xiao et al., 2026). Regarding Encoder-Only models, we selected ModernBERT (Warner et al., 2025), DeBERTa (He et al., 2020), and RoBERTa (Liu et al., 2019). For Decoder-Only models, we opted for the Qwen3 series and the Llama 3.1 series. Additionally, we employed DeepSeek V3.1 to perform CoT distillation.

Hyper-Parameter Selection. Regarding the API configuration, the temperature was set to 0 to ensure reproducibility, and the reasoning mode was enabled. For the finetuning approach, we used the Hugging Face trl [2]library for SFT, in conjunction with the unsloth [3] library to accelerate training. Regarding the specific training hyperparameters, the learning rate was set to 2e-4 and the number of epochs to 3. Furthermore, we employed 4-bit quantization (Banner et al., 2019) and integrated LoRA (Hu et al., 2022), setting the rank of the low-rank matrices to 16.

Evaluation Metrics. The evaluation metric is the macro F1-score. To provide a more comprehensive assessment of model performance, we additionally include Accuracy as a supplementary metric.

[1] https://huggingface.co/datasets/ailsntua/QEvasion

[2] https://github.com/huggingface/trl

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

Figure 3: Confusion matrix of the Llama-3.1-8B model fine-tuned via CoT distillation on Task 2.

5 Results and Analysis

The final experimental results are presented in Table 1. Among Encoder-Only models, ModernBERT outperforms other BERT variants by a significant margin. Its peak performance approaches that of SFT-adapted Decoder-Only models. This demonstrates the superiority of the latest BERT architecture on this task. Within Decoder-Only models, the best performance is achieved by the Llama model trained with CoT Distillation. Performance gains from distillation are clearly evidenced by comparing non-distilled and distilled models. Finally, the highest F1 scores for both Task 1 and Task 2 are obtained by MiMo-V2-Flash utilizing HTP. This underscores the robust capability of the proposed HTP. The strategy effectively guides the model to reason and subsequently enhances performance.

Prompt Engineering. In this paper, multiple fundamental prompting strategies are employed to benchmark against the proposed HTP. For one-shot and few-shot, analysis processes generated by DeepSeek are incorporated into the prompt. This inclusion assists the model in better comprehending the analytical procedure. For few-shot, a single example is added for Clear Reply, Clear Non-Reply, and Ambivalent Reply respectively. Regarding CoT, the model is instructed to explicitly generate the analysis process before the final label. The quantitative results are presented in Table 2. HTP achieves superior performance across all metrics. As indicated by the data, performance on Task 2 improves as the number of examples increases.

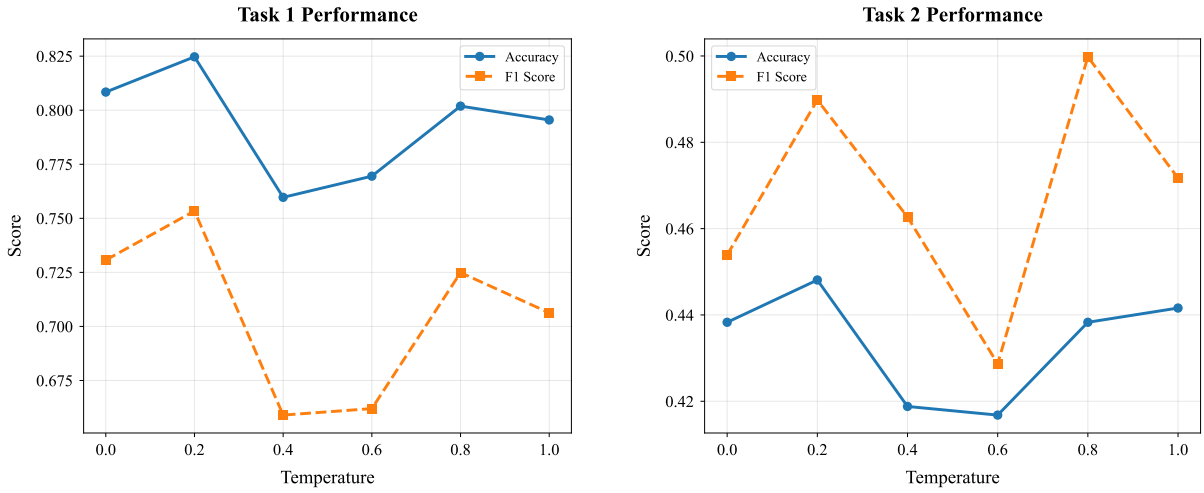

Figure 4: Performance of the Llama-3.1-8B model on Task 1 and Task 2 at different temperature settings.

Conversely, performance on Task 1 remains largely unaffected.

Error Analysis. To further investigate the performance bottlenecks on the fine-grained evasion classification in Task 2, we analyzed the confusion matrix of the predictions, as illustrated in Figure 3. The distribution reveals that the misclassifications are not random but exhibit distinct, concentrated biases.

The most prominent observation is the model's overwhelming tendency to classify various distinct rhetorical strategies as Deflection. For instance, a substantial number of samples are misclassified into the Deflection category, specifically 35 Explicit instances, 37 Implicit instances, and 34 Dodging instances. This indicates that while the CoT-distilled model successfully detects the overarching presence of an evasive maneuver, it frequently adopts Deflection as a safe, default label when confronting ambiguous semantics.

The matrix highlights the severe difficulty in separating Dodging from Deflection. Although the model correctly identifies Dodging in 11 instances, the high misclassification rate of 34 instances shows that the linguistic bridges politicians use to talk around a question often mimic direct topic deflection. Furthermore, Implicit replies are rarely classified correctly, with only 11 instances identified accurately, and are heavily confused with Deflection. This suggests that the student model occasionally lacks the cultural or contextual world knowledge required to decode implied meanings, thus failing the logic check.

Temperature Sensitivity. The temperature parameter was tuned to enhance the performance of Decoder-Only models. During the experiments, top_p was fixed at 1 and max_new_tokens was set to 4096. Fine-grained tuning was conducted on the Llama-3.1-8B model. Specifically, the temperature was increased from 0 to 1 in increments of 0.2. As illustrated in Figure 4, the poorest performance is observed at intermediate temperatures and the optimal performance is achieved at a temperature of 0.8.

6 Conclusion

In this paper, we presented the YNU-HPCC system for SemEval-2026 Task 6. We focused on the challenge of classifying clarity and evasion in political interviews. To address this, we proposed two methods: Hierarchical Taxonomy Prompting and CoT distillation. HTP guides the model to identify the clarity level first, then the specific evasion strategy. CoT distillation transfers reasoning capabilities from the DeepSeek teacher model to smaller student models. Experiments show that our approach is effective. The system achieved 6th place in Task 1 and 5th place in Task 2.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Ron Banner, Yury Nahshan, and Daniel Soudry. 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in neural information processing systems*, 32.

Kaiyuan Chen, Jin Wang, and Xuejie Zhang. 2024. Mathematical reasoning via multi-step self questioning and answering for small language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 81–93. Springer.

Kaiyuan Chen, Jin Wang, and Xuejie Zhang. 2025. Learning to reason via self-iterative process feedback for small language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3027–3042.

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: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tiesunlong Shen, Rui Mao, Jin Wang, Heming Sun, Jian Zhang, Xuejie Zhang, and Erik Cambria. 2026. Llmdoctor: Token-level flow-guided preference optimization for efficient test-time alignment of large language models. *arXiv preprint arXiv:2601.10416*.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. " i never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. Semeval-2026 task 6: Clarity – unmasking political question evasions. *Preprint*, arXiv:2603.14027.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, and 1 others. 2026. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*.

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*.

Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. 2024. Enhancing semantics in multimodal chain of thought via soft negative sampling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6059–6076.