AGAI at SemEval-2026 Task 10: Enhancing Conspiracy Detection via Instruction-tuned LLMs

Anonymous ACL submission

Abstract

This paper presents our solution for subtask2, which focuses on the automated detection of conspiracy in text. Unlike traditional toxic text detection, conspiracy identification is particularly challenging as it often lacks explicit semantic indicators. To address this, we leveraged a Large Language Model (LLM) as our backbone and employed Low-Rank Adaptation (LoRA) for fine-tuning to enhance detection performance. To generate probabilistic confidence scores while maintaining the generative capabilities of the LLM, we implemented a hybrid loss function that integrates both generative and token classification losses. Additionally, semi-supervised learning with unlabeled data was incorporated to further refine classification accuracy. Our approach achieved a test accuracy of 0.87, ranking 2nd in both stages of the competition leaderboard.

1 Introduction

Psycholinguistic conspiracy(PsyCo) have transcended the fringes of the internet to become a significant sociopolitical force, often fueling misinformation and eroding public trust. Recent psychological research suggests that conspiracy thinking is not merely about what is said, but how it is structured, characterized by specific cognitive patterns(Liu et al., 2024). Subtask 2 of SemEval-2026 Task 10 provides a high-quality dataset designed to benchmark the automatic detection of PsyCo in text. In SemEval-2026 task 10, unlike prior datasets that are tied to specific events, PsyCoMark is grounded in the structural markers of conspiratorial thought. This task shifts the focus from superficial keywords to the underlying psycholinguistic architecture of a claim. To this end, PsyCoMark utilize a topic-diverse dataset sourced from real-world Reddit discussions, capturing the messy and organic nature of everyday online discourse.

To effectively identify PsyCo in text, this paper proposes a framework based on Large Language Models (LLMs). While traditional text classification models, such as the BERT family(Devlin et al., 2019), capture semantic representations through pre-training, they are often constrained by model scale and architectural limitations. Consequently, they struggle to achieve a profound understanding of the complex world knowledge and nuanced reasoning. In contrast to the BERT framework, which primarily functions as a bidirectional encoder requiring extensive task-specific fine-tuning, LLMs exhibit a paradigm shift in both generalization and semantic depth. By leveraging unprecedented parameter scales and diverse training corpora, LLMs transcend the limitations of discriminative models, demonstrating superior zero-shot and few-shot learning capabilities that allow them to generalize across unseen domains without additional parameter updates(OpenAI et al., 2024). Furthermore, while BERT excels at token-level representations, LLMs manifest emergent abilities in complex semantic understanding and logical reasoning, enabling a more nuanced grasp of human intent and long-range contextual coherence.

We propose a training framework for psycholinguistic conspiracy detection via domain-specific fine-tuning of LLMs. Our primary contributions consist of three parts:

1. Semi-supervised Learning Integration: We exploit unlabeled data through a semi-supervised approach, significantly boosting the model's precision in identifying conspiracy-related content.

2. Generative-Discriminative Fusion: We leverage the generative priors of LLMs by integrating generative objectives with classification scoring. This multi-task learning strategy, stabilized by a fused loss function, enhances overall classification performance.

3. Establishing a robust and adaptable architecture that generalizes effectively across various thematic domains.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2065–2069
July 3-4, 2026 ©2026 Association for Computational Linguistics

2 Our Approach

2.1 Framework

Figure 1 illustrates our end-to-end algorithmic framework for the SemEval competition. Our methodology begins with data augmentation: to utilize the unlabeled portion of the training set, we employ a semi-supervised learning strategy, assigning soft labels via a fine-tuned LLM to facilitate downstream training. The core architecture involves fine-tuning the Qwen3-14B model using LoRA (Low-Rank Adaptation) (Hu et al., 2021). Our optimization strategy aims to preserve the model's inherent generative reasoning capabilities while simultaneously refining its discriminative precision. To this end, the objective function is a weighted combination of three components: (1) a standard generative loss to maintain linguistic proficiency(Chu et al., 2025); (2) a cross-entropy classification loss applied to the target tokens (e.g., "yes"/"no") for binary discrimination; and (3) a Kullback-Leibler (KL) divergence loss(Hinton et al., 2015) to align the model with the soft labels.

2.2 Training Strategy

For the data preprocessing part, the labeled dataset is partitioned into five subsets for five-fold cross-validation and soft label generation. For the LLM training stage via LoRA (Low-Rank Adaptation), we employ a composite loss function consisting of a weighted average of generative loss, token-level cross-entropy loss, and KL divergence. Example of a token-level cross-entropy loss is as follows,

$$\mathcal{L}_{tc} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{1}$$

where y_i represents the ground truth label, taking values of 0 or 1. Where $\hat{y}_i$ denotes probability for the target token by the model output. The generative loss is defined in Eq. 2,

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x) \tag{2}$$

where T represents the length of the sequence after removing prompt tokens. The $P(y_t \mid y_{<t}, x)$ denotes the conditional probability distribution of predicting the target token y_t given the preceding context $y_{<t}$. An example of a KL divergence loss is defined in Eq. 3,

$$\mathcal{L}_{KL} = \sum_{i=1}^{V} p_i^T \log\left(\frac{p_i^T}{p_i^S}\right) \tag{3}$$

where $\mathcal{L}_{KL}$ reference the loss function of standard knowledge distillation. p_i^T represents the soft label was predicted from the tuned-llm model. p_i^S denotes the model's prediction probability distribution in the training stage. The final training loss such as Eq. 4

$$\mathcal{L} = \lambda_1\mathcal{L}_{tc} + \lambda_2\mathcal{L}_{gen} + \lambda_3\mathcal{L}_{KL} \tag{4}$$

where λ_1, λ_2, and λ_3 represent hyperparameters.

The implementation details of the three loss functions during the training process are as follows. First, the generative loss adopts the standard objective used in large language model fine-tuning, where "Yes" and "No" are defined as the target tokens in the output sequence. Second, the token-level classification loss calculates the cross-entropy between the predicted key tokens and their corresponding ground truth labels. Specifically, "Yes" and "No" are mapped to 1 and 0 to facilitate binary classification. Regarding the KL divergence loss, we employ a five-fold cross-validation approach utilizing the aforementioned two loss functions to generate soft labels for all training samples. For unlabeled data, the average prediction from the five-fold models serves as the soft label(Kingma et al., 2014). Finally, the model is fine-tuned by integrating all three loss functions to produce the definitive inference model.

3 Experiments

3.1 Experimental Settings

Dataset The experimental dataset was constructed based on the official competition training set, with entries labeled 'can't tell' explicitly excluded. We conducted five-fold cross-validation for offline experiments focusing on the 'no' and 'yes' labels. In each fold, four-fifths of the data were utilized for training, while the remaining portion served as the validation set. Notably, the previously excluded 'can't tell' data were re-integrated via soft label augmentation to enhance the training process, ensuring no data leakage into the validation set.

Metric To comprehensively evaluate the performance of our model across all categories, competition employ the Macro-averaged F1-score as our primary evaluation metric. Unlike the micro-average, which can be dominated by majority classes, the macro-average treats all classes equally by calculating the metric independently for each class and then taking the average. This ensures that

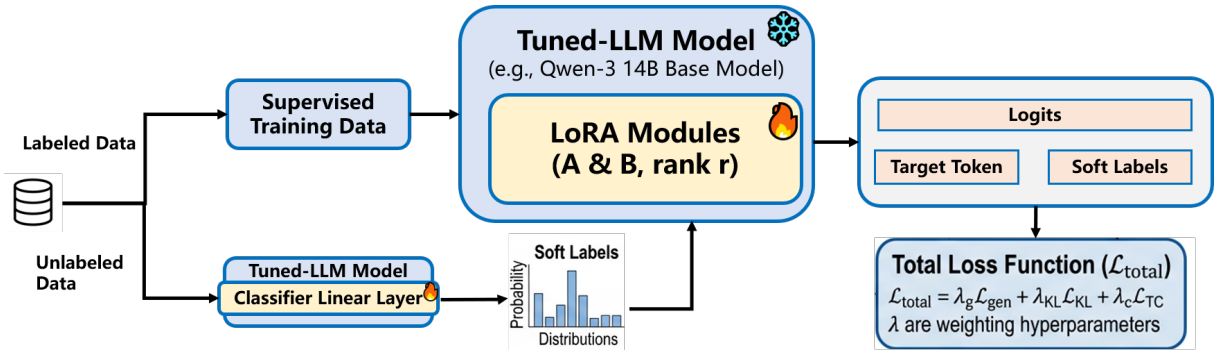

Figure 1: Conspiracy Detection LLM Training Framework.

Fine-tuning Model	Val F1
`Qwen2.5-7B`	79.28 ± 1.25
`Qwen2.5-14B`	82.92 ± 0.78
`Qwen3-14B`	**84.28 ± 1.60**
`Gemma-2-9b`	83.10 ± 1.25

Table 1: F1-scores of various LLMs on local five-fold cross-validation (percentage values). The reported metrics represent the average F1-score across five folds, with fluctuations indicated based on five random runs.

the performance on minority classes is adequately reflected.

Experimental Setup All models were trained on a high-performance computing cluster equipped with NVIDIA A100 GPUs. The proposed model is implemented using the PyTorch framework. For the training of LLMs, the Huggingface Transformers library(Wolf et al., 2020) is utilized, with all pre-trained weights sourced via the Huggingface Hub API

Model Input This paper uses a consistent prompt, with the fixed prompt word text appended to the beginning of the original input text. The prompt prefix is **"Please serve as a user review expert, detect whether a Reddit comment expresses a conspiracy belief"**.

3.2 Experimental Results

In this study, we evaluated the performance of various fine-tuned LLMs. The experimental results, detailed in Table 1, were obtained using open-source models, including the Qwen(Yang et al., 2025) and Gemma(Team et al., 2024) series.

Empirical results from Table 1 indicate that model performance scales positively with both size and recency. These findings underscore the critical role of the base model, suggesting that larger architectures inherently yield superior outcomes following fine-tuned.

We also conducted ablation experiments on different modules in our framework. The ablation implementation using qwen3-14B as the backbone is shown in Table 2. Experimental results from Table 2 show that different modules all have a positive impact on the final result. Directly using the generation loss function to construct question-answer pairs and calculate the error yields the accuracy result. Looking at the fluctuations in the results, introducing the classification cross-entropy loss function improves the result while also making the cross-entropy smaller. Further utilizing semi-supervised learning to introduce soft labels and calculate KL divergence improves the overall result, but also makes it larger. Currently, only a simple weighted combination loss function is used; future designs could consider more consistent approaches to improve the stability of the results.

Furthermore, we conducted comparative experiments to evaluate the impact of various LoRA configurations across different pre-trained models, the results of which are summarized in Table 3. Experimental results demonstrate that optimal performance for various pre-trained models contingent upon specific LoRA parameter configurations. Among all tested architectures, the qwen3-14B model, configured with a rank of 68 and an alpha of 128, achieved the superior performance.

4 Conclusion

This paper investigates the task of conspiracy Detection within textual data using LLMs. By integrating token-level classification loss into a generative framework, we effectively bridge the gap between generative pre-training and discriminative tasks. Furthermore, we employ a semi-supervised knowl-

Lora model	Generative Loss	Token-level cross-entropy loss	KL Loss	Val F1 score
•	•			78.36 ± 1.40
•		•		80.67 ± 0.76
•	•	•		83.12 ± 1.05
•	•		•	80.56 ± 1.34
•		•	•	83.78 ± 1.24
•	•	•	•	**84.28 ± 1.60**

Table 2: The table shows ablation experiments with different module combinations within the framework.

Model	Rank	Alpha	Avg F1
Gemma-2-9b	8	16	78.70
	16	32	82.28
	32	**64**	**83.10**
	64	128	82.65
Qwen2.5-7b	8	16	77.58
	16	32	78.20
	32	**64**	**79.28**
	64	128	77.74
Qwen2.5-14B	8	16	81.10
	16	32	79.88
	32	64	78.56
	64	**128**	**82.92**
Qwen3-14B	8	16	80.62
	16	32	83.74
	32	64	82.86
	64	**128**	**84.28**

Table 3: Comparison of different LoRA hyperparameter configurations across various LLMs

edge distillation strategy, utilizing LLM-generated soft labels for unlabeled data to maximize data utility and boost classification robustness. While our current approach focuses on algorithmic modeling and data augmentation, future research could benefit from incorporating domain-specific insights, such as psychological knowledge, to refine detection in nuanced scenarios. Additionally, it is worth exploring techniques that enhance classification accuracy primarily through inference-time optimization rather than fine-tuning.

Acknowledgments

We thank the organizers of SemEval-2026 Task 10: Psycholinguistic Conspiracy Marker Extraction and Detection for creating a well-structured task

References

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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

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.

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. 2014. Semi-supervised learning with deep generative models. *Preprint*, arXiv:1406.5298.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. *ConspEmoLLM: Conspiracy Theory Detection Using an Emotion-Based Large Language Model*. IOS Press.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.