Team Evaluators at SemEval-2026 Task 6: Instruction-Tuned LLMs for Clarity and Evasion Classification in Political Interviews

Siva Gopala Krishna Nuthakki[1], **Sanjay Reddy**[1], **Sai Tejaswi Woona**[1]
[1]BML Munjal University, Haryana, India

Abstract

This work is part of the SemEval-2026 CLARITY shared task (Task 6), which focuses on detecting clarity and evasion in political question–answer pairs from interviews and debates. The competition includes two subtasks: clarity-level classification (Clear Reply, Ambiguous, Clear Non-Reply) and evasion-level classification, which identifies one of nine fine-grained evasion techniques. The dataset consists of annotated question–answer pairs with hierarchical labels for both clarity and evasion, enabling comprehensive evaluation of nuanced discourse phenomena. We fine-tune open-source large language models using Low-Rank Adaptation (LoRA) and supervised fine-tuning (SFT), employing structured prompts that jointly encode the question and answer to capture discourse cues. Models are evaluated using Macro F1, the official metric of the shared task. Our system achieves a Macro F1 of 0.83 on Subtask 1 (5th place) and 0.54 on Subtask 2 (9th place), demonstrating that parameter-efficient fine-tuning of LLMs is effective for modeling strategic ambiguity in political discourse.

1 Introduction

Political communication often relies on strategic ambiguity, where speakers avoid directly answering questions while maintaining plausible deniability. In high-stakes settings, such as televised debates and interviews, politicians often employ evasive discourse techniques that obscure intent, shift focus, or only partially address the question. Although equivocation has been extensively studied in political science, its automatic detection remains challenging in natural language processing because of subtle semantic cues, pragmatic reasoning, and contextual dependencies.

The SemEval-2026 CLARITY shared task (Task 6) (Thomas et al., 2026) advances computational research in this direction by introducing a hierarchical framework for analyzing response clarity in political question–answer (QA) pairs. The task defines two related subtasks:

1) **Task 1 – Clarity-level Classification:** Determine whether a response is Clear Reply, Ambiguous, or Clear Non-Reply with respect to the question.

2) **Task 2 – Evasion-level Classification:** Classify ambiguous or evasive responses into one of nine fine-grained evasion techniques derived from political discourse theory.

Both subtasks are formulated as multi-class classification problems and evaluated using Macro F1 to ensure balanced performance across categories. The hierarchical labeling scheme captures both high-level clarity distinctions and nuanced rhetorical strategies, making the task challenging due to overlapping semantic signals, implicit reasoning, and discourse-level phenomena.

In this work, we participate in both subtasks and focus on developing an efficient and scalable approach using instruction-tuned large language models. We fine-tune open-source LLMs through Low-Rank Adaptation (LoRA)(Hu et al., 2021) combined with supervised fine-tuning (SFT), enabling parameter-efficient adaptation without updating all model weights. Our approach is designed to model strategic ambiguity effectively while delivering strong performance in both clarity and evasion classification tasks.

2 Methodology

2.1 Dataset Preparation

We utilize the QEvasion dataset (Thomas et al., 2024) introduced in *"I Never Said That": A Dataset, Taxonomy and Baselines on Response Clarity Classification*, which serves as the official benchmark for the SemEval-2026 CLARITY shared task. The dataset consists of question–answer (QA) pairs extracted from political interviews and annotated according to a two-level hier-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3442–3445

archical taxonomy of response clarity and evasion. Each instance corresponds to a sub-question derived from an interview question and the associated interviewee response, enabling fine-grained evaluation of how effectively the answer addresses a specific query.

The dataset is divided into predefined training and test splits, containing 3,448 and 308 instances, respectively. Each example includes metadata such as interview title, date, president (when available), and source URL, along with the core fields *question* and *interview answer*. Annotations are provided at two levels: (1) a clarity label representing the top-level categories Clear Reply, Clear Non-Reply, and Ambivalent Reply and (2) an evasion label corresponding to one of nine fine-grained subcategories, including Explicit, Implicit, Dodging, General, Deflection, Partial/half-answer, Declining to answer, Claims ignorance, and Clarification. Table 1 reports the Task 1 clarity-label distribution (Codabench, a), and Table 2 reports the Task 2 evasion-label distribution(Codabench, b).

Annotation Type	Training Set
Ambivalent	2040
Clear Non-Reply	356
Clear Reply	1052

Table 1: Task 1: Clarity label distribution on the training set.

Annotation Type	Training Set
Claims ignorance	119
Clarification	92
Declining to answer	145
Deflection	381
Dodging	706
Explicit	1052
General	386
Implicit	488
Partial/half-answer	79

Table 2: Evasion label distribution (Task 2) on the training set.

To prepare the data for supervised fine-tuning (SFT), we reformatted each instance into an instruction-based structure. Each training sample consists of: (1) an instruction prompt defining the classification task and label definitions, (2) the input composed of the interview question and corresponding answer, and (3) the target output containing the gold clarity or evasion label. This structured format enables the model to jointly reason about the intent of the question and the informational adequacy of the response. For Subtask 1, the output space is restricted to the three clarity categories, while for Subtask 2, the model predicts one of the nine evasion techniques. The consistent representation facilitates parameter-efficient adaptation of large language models using LoRA-based fine-tuning across both tasks.

2.2 Model Overview

We propose two complementary approaches: (1) Supervised Fine-Tuning (SFT) using parameter-efficient LoRA adapters, and (2) a hybrid approach (SFT + few-shot prompting) that integrates an additional verification stage during inference. Both methods aim to improve the model's ability to classify response clarity and identify fine-grained evasion strategies in political question–answer pairs.

We employ three instruction-tuned large language models: Phi-4-14B-Instruct, Qwen3-14B, and Ministral-3-7B. These models are selected for their strong reasoning capabilities and robustness in instruction-following tasks, which are essential for discourse-level semantic classification. To enable efficient adaptation, we apply LoRA adapters (rank = 8, α = 32, dropout = 0.1) to the attention and feed-forward layers, allowing task-specific updates while keeping the majority of pretrained parameters frozen.

The task is framed as generative classification, where the model outputs one of the predefined categorical labels. For Subtask 1, the model predicts one of three clarity labels, while for Subtask 2, it predicts one of nine fine-grained evasion techniques. A detailed explanation of our two approaches is given below.

1) Supervised Fine-Tuning (SFT). The model is fine-tuned using cross-entropy loss with the AdamW optimizer (learning rate = 2e–4, training for 3 epochs, gradient accumulation = 8). To handle class imbalance, we apply weighted loss. Separate models are trained for each subtask to better capture task-specific distinctions.

2) Hybrid Approach (SFT + Few-shot Prompting). The hybrid method builds upon the SFT model by incorporating an LLM-based verification stage. During inference, the SFT-generated predictions are provided to a large language model, which is prompted to assess whether the predicted labels are correct based on the conversation context.

3 Results and Analysis

This section presents the performance of our systems on both subtasks of the CLARITY shared task. All models are evaluated using Macro F1-score, the official metric, ensuring balanced performance across classes. We report results on both the development phase and the final official test phase.

3.1 A. Subtask 1: Clarity-level Classification

Table 3 summarizes the results for Subtask 1. On the development set, Phi-4-14B-Instruct achieved the highest performance (Macro F1 $= 0.72$), outperforming Qwen3-7B (0.63) and the hybrid approach (0.43).

On the official test set, Phi-4-14B-Instruct further improved to a Macro F1 of 0.83, ranking 5th overall among participating teams. Qwen3-14B also performed competitively with a Macro F1 of 0.80. In contrast, the hybrid verification pipeline did not yield improvements and showed significantly lower stability during development.

Model / Approach	Dev F1	Test F1
Phi-4-14B-Instruct	0.72	0.83
Qwen3-7B	0.63	–
Qwen3-14B	–	0.80
Hybrid (Qwen3 + LLaMA)	0.43	–

Table 3: Subtask 1 (Clarity Classification) results – Macro F1.

The results indicate that clarity-level classification benefits strongly from larger instruction-tuned models fine-tuned with LoRA. The substantial gap between SFT models and the hybrid pipeline suggests that additional verification may introduce noise rather than improving decision boundaries. Errors primarily occur in borderline cases between Ambivalent Reply and Clear Reply, where pragmatic interpretation plays a central role.

B. Subtask 2: Evasion-level Classification

Subtask 2 presents a significantly more challenging setting due to the nine-class fine-grained taxonomy. As shown in Table 4, Phi-4-14B-Instruct achieved the best development performance (Macro F1 = 0.56), followed by Ministral-3 (0.51) and the hybrid approach (0.49).

On the official test set, Phi-4-14B achieved a Macro F1 of 0.54, ranking 9th overall. In contrast, the hybrid model dropped substantially to 0.31, indicating poor generalization in fine-grained evasion prediction.

Model / Approach	Dev F1	Test F1
Phi-4-14B-Instruct	0.56	0.54
Ministral-3-7B	0.51	–
Hybrid (Qwen3 + LLaMA)	0.49	0.31

Table 4: Subtask 2 (Evasion Classification) results – Macro F1.

Compared to Subtask 1, performance drops notably due to the increased label granularity and semantic overlap among evasion categories (e.g., Deflection vs. Dodging, Implicit vs. General). Rare categories further affect Macro F1, making balanced prediction more difficult. The hybrid approach suffers from overcorrection in ambiguous cases, leading to unstable predictions and reduced generalization on the test set.

C. Key Observations

Across both subtasks, LoRA-based Supervised Fine-Tuning consistently outperforms the hybrid verification approach. Larger 14B-parameter models demonstrate stronger discourse reasoning capabilities, particularly in clarity-level classification. While hybrid verification can occasionally correct obvious misclassifications, it introduces instability and harms performance on fine-grained evasion categories.

Overall, our findings show that carefully tuned, parameter-efficient SFT models are highly effective for modeling strategic ambiguity in political discourse, especially when combined with structured instruction design and constrained generative outputs.

4 Conclusion

This study demonstrates that parameter-efficient fine-tuning of instruction-tuned large language models can effectively model response clarity and fine-grained evasion strategies in political discourse. By adapting Phi-4-14B-Instruct, Qwen3-14B-Thinking, and Ministral-3-7B using LoRA-based Supervised Fine-Tuning (SFT), our approach achieved competitive performance on both subtasks of the SemEval-2026 CLARITY shared task. In particular, our system attained a Macro F1 of 0.83 on Clarity-level Classification (Subtask 1) and 0.54 on Evasion-level Classification (Subtask 2), securing strong leaderboard rankings.

Our experiments show that structured instruction design and parameter-efficient adaptation are more effective than hybrid verification pipelines for hierarchical discourse classification. While the hybrid

approach occasionally corrected ambiguous cases, it introduced instability and reduced generalization performance, especially in fine-grained evasion prediction. Larger 14B-parameter models consistently outperformed smaller architectures, highlighting the importance of reasoning capacity in modeling pragmatic and discourse-level phenomena.

Overall, this work provides evidence that open-source LLMs, when enhanced through lightweight LoRA-based fine-tuning and carefully designed instruction prompts, can reliably detect strategic ambiguity in political communication. These findings contribute to the development of scalable tools for political discourse analysis and offer promising directions for future research in explainable and robust ambiguity detection systems.

5 References

References

Codabench. a. SemEval 2026 Task 6: CLARITY – Subtask 1: Clarity Level. Codabench Competition 10879, [Online]. [Online]. Available: https://www.codabench.org/competitions/10879/. [Accessed: Mar. 03, 2026].

Codabench. b. SemEval 2026 Task 6: CLARITY – Subtask 2: Evasion Level. Codabench Competition 11131, [Online]. [Online]. Available: https://www.codabench.org/competitions/11131/. [Accessed: Mar. 03, 2026].

E. Hu, Y. Shen, P. Wallis, and 1 others. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

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, Miami, Florida, USA. Association for Computational Linguistics.

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.