The Argonauts at SemEval-2026 Task 6: Large Language Models for Response Clarity Classification: Prompting, Fine-Tuning, and Data-Centric Approaches

Sajib Bhattacharjee*, **Sha Newaz Mahmud***, **Md. Refaj Hossan, Kawsar Ahmed and Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
u2004003@student.cuet.ac.bd, moshiul_240@cuet.ac.bd

Abstract

Detecting equivocation is essential, as indirect or evasive responses can shape public perception, influence political narratives, and undermine transparency in democratic discourse. To address the challenge of detecting evasive political responses on digital platforms, participation in the CLARITY SemEval-2026 Task was undertaken, which focuses on (i) clarity-level classification and (ii) fine-grained evasion-type classification in political question-answer contexts. This study introduces a data-centric framework that systematically examines the effects of class distribution and refinement strategies on the performance of Large Language Models (LLMs). A distribution-aware, LLM-augmented dataset was constructed by selectively paraphrasing minority-class instances to enhance class balance, and its performance was benchmarked against full, rebalanced, and undersampled training configurations. To comprehensively assess the proposed method, Qwen3-14B, Phi-4, Gemma-2 9B, and Mistral 7B were evaluated in in-context learning (ICL) settings (zero-shot and few-shot) and with LoRA fine-tuning. Experimental results indicate that fine-tuning Phi-4 with class rebalancing yields strong performance, achieving 74.77% on Subtask-1 and 51.55% on Subtask-2. Consequently, the system ranked 21^{st} in Subtask-1 and 22^{nd} in Subtask-2 on the official evaluation leaderboard.

1 Introduction

Political communication is inherently strategic, especially in high-stakes settings such as televised debates and presidential interviews, where speakers often give indirect or ambiguous responses. Instead of directly answering questions, politicians may shift topics, use vague language, or repeat prepared talking points. This strategy is known as equivocation or evasion, used to manage public perception and mitigate risk (Thomas et al., 2024). In today's fast-paced media landscape, the strategic use of ambiguity has even greater impact, as televised and social media messages rapidly shape public narratives. Evasive responses can influence how audiences interpret political positions, often leaving voters uncertain about leaders true stances. Despite extensive research in political science, automatic evasion detection remains under-explored in computational linguistics.

To address these challenges, we participated in the CLARITY shared task (Thomas et al., 2026) at SemEval-2026 (Ghosh et al., 2026),[1] which comprises two subtasks: (i) clarity-level classification of answers and (ii) fine-grained evasion-type classification, which focuses on detecting and analyzing question evasion and answer clarity in English political discourse. To solve this task, this work constructs a distribution-aware LLM-augmented dataset by selectively paraphrasing minority-class and systematically compares it against the original, rebalanced, and undersampled dataset configurations. The main contributions of this work are as follows:

- Propose a data-centric training framework that utilizes LLM-based refinement to rebalance the CLARITY dataset and systematically compares it with original, rebalanced, and undersampled variants.
- Provide a comprehensive empirical study of four decoder-only LLMs across ICL and LoRA fine-tuning settings, offering detailed insights into their behavior under low-resource and balanced-data scenarios.
- Demonstrate that performance gains stem primarily from controlled refinement and class-distribution strategies, rather than from

*Authors contributed equally to this work.

[1] https://konstantinosftw.github.io/CLARITY-SemEval-2026/

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2729–2743
July 3-4, 2026 ©2026 Association for Computational Linguistics

model scale alone, showing that carefully designed data setups enable mid-sized LLMs to deliver competitive results.

Several challenges persist in the current implementation. Performance remains sensitive to class imbalance, particularly for low-frequency evasion categories, while paraphrasing with large language models can introduce distribution shifts. Additionally, most models fail to capture task-specific structure in zero-shot and few-shot scenarios, resulting in limited task alignment. These limitations underscore the necessity for improved data strategies and enhanced generalization across training methodologies.

2 Literature Review

Equivocation, or evasion in political discourse, has been extensively studied in linguistics and political science prior to computational approaches. Dillon (2025) characterized it as a strategy for responding without directly answering a question. Harris (1991) and Bull and Mayer (1993) analyzed political interviews, distinguishing replies from non-replies and identifying systematic evasion patterns. Bull (2003) proposed a structured framework distinguishing direct replies, indirect replies, and various forms of non-replies.

Despite extensive social science research, equivocation has only recently gained attention in computational linguistics. Earlier NLP work focused on related topics, such as discourse structure and conversational intent, rather than on response clarity itself. Majumder et al. (2020) constructed a large-scale political dialogue dataset and modeled persuasion and discourse strategies. Contextualized representations of political agendas in social media discourse have been examined by Pujari and Goldwasser (2021). Another adjacent field concerns deception detection. Girlea (2017) studied linguistic cues of deceptive intent using probabilistic models, and Ferracane et al. (2021) introduced a dataset to analyze whether political speakers answered questions and how sincere their responses were perceived to be. Although NLP studies political discourse at scale, equivocation, as a clarity classification task, remains underexplored. Trotta and Tonelli (2021) examined multimodal aspects of political interviews, including gestures and non-verbal cues. In contrast, we treat response clarity as a structured classification problem using the dataset introduced by Thomas et al. (2024).

In contrast to prior work's theoretical analyses or adjacent NLP tasks (persuasion, deception, discourse), our approach adopts a data-centric perspective, i.e., we use the dataset to systematically investigate how controlled LLM-based refinement, class rebalancing, and training strategies influence performance on clarity and evasion classification, emphasizing thoughtful data engineering over model scale to achieve scalable outcomes.

3 Dataset and Task Description

The CLARITY SemEval-2026 shared task comprises two subtasks focused on classifying answer clarity levels and fine-grained evasion types in political discourse.

- **Subtask 1 (Clarity-level Classification):** This subtask[2] requires classifying an answer, given a question-answer pair, into one of three predefined categories.
- **Subtask 2 (Evasion-level Classification):** The second subtask[3] focuses on fine-grained evasion detection. Given a question-answer pair, the goal is to classify the answer into one of nine predefined evasion techniques.

Both subtasks draw on the dataset introduced by Thomas et al. (2024), which provides a hierarchical taxonomy and benchmark for response clarity and evasion classification in political QA settings. Table 1 shows that the dataset comprises 3,348 training instances, 308 development instances, and 237 test instances, along with total words, total unique words, and average sample length across various dataset settings.

Datasets	T_S	T_W	T_{UW}	L_{Avg}
Original Dataset	3,348	1,061,074	24,006	307.74
Rebalanced Dataset	2,250	620,391	18,289	275.73
Refined Dataset	3,101	921,627	23,329	297.2
Undersampled Dataset	711	193,581	12,156	272.27
Development Dataset	308	105,100	5,864	341.23
Test Dataset	237	34,748	3,510	146.62

Table 1: Counts of total samples (T_S), total words (T_W), unique words (T_{UW}), and average sample length (L_{Avg}) across all dataset settings.

4 System Overview

This section outlines the data-centric approach to the task, including the problem formulation, data refinement strategy, model configurations, and the overall pipeline.

[2] https://www.codabench.org/competitions/10879/

[3] https://www.codabench.org/competitions/11131/

4.1 Problem Formulation

The CLARITY shared task is formulated as a hierarchical multi-class text classification problem over political QA pairs.

Let $\mathcal{D} = \{(q_i, a_i, y_i^c, y_i^e)\}_{i=1}^N$ denote the labeled dataset, where q_i represents the question and a_i represents the corresponding answer. For Subtask 1, the clarity label y_i^c belongs to one of three classes: *Clear Reply*, *Ambivalent*, or *Clear Non-Reply*. For Subtask 2, the fine-grained evasion label y_i^e belongs to one of nine categories: *Explicit*, *Implicit*, *Dodging*, *General*, *Deflection*, *Partial*, *Declining*, *Ignorance*, or *Clarification*.

Following the taxonomy proposed by Thomas et al. (2024), a deterministic mapping from fine-grained evasion types to high-level clarity categories is defined as follows: *the Explicit type maps to Clear Reply; the Implicit, Dodging, General, Deflection, and Partial types all map to Ambivalent; and the Declining, Ignorance, and Clarification types all map to Clear Non-Reply.*We model both subtasks jointly using a generation-based formulation. Given an input prompt constructed from the question-answer pair $x_i = (q_i, a_i)$, let z denote the generated target label sequence enclosed within the predefined <LABEL> tags, corresponding to one of the fine-grained evasion categories. Prediction is obtained as shown in Eq. (1).

$$\hat{z}_i = \arg\max_z P(z \mid x_i; \theta), \tag{1}$$

from which the fine-grained evasion label is extracted as defined in Eq. (2).

$$\hat{y}_i^e = \text{Parse}(\hat{z}_i), \qquad \hat{y}_i^c = f(\hat{y}_i^e). \tag{2}$$

Supervised fine-tuning minimizes the token-level cross-entropy (negative log-likelihood) of the target label sequence z_i conditioned on the input x_i, as shown in Eq. (3).

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{|z_i|} \log P(z_{i,t} \mid x_i, z_{i,<t}; \theta)\,. \tag{3}$$

Figure 1 presents the overall system architecture. The implementation, source code, and the augmented dataset are publicly available on GitHub.[4]

[4] https://github.com/Sojib001/CLARITY-Political-Question-Evasions

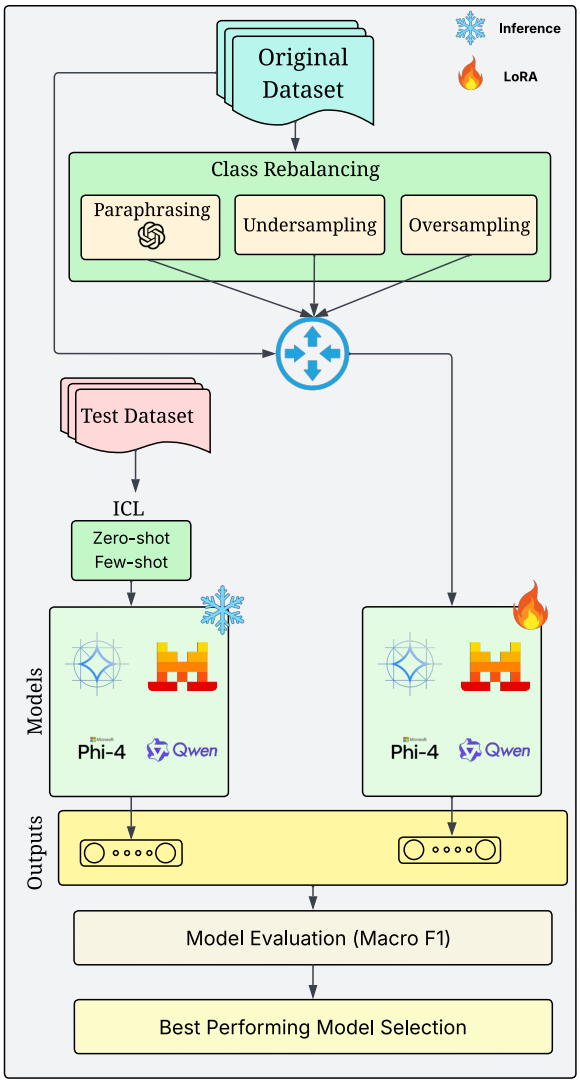


Figure 1: Abstract representation of our methodology pipeline, including data rebalance, LLM-based approaches, and model evaluation.

4.2 Class rebalancing

The original dataset exhibits notable class imbalance across clarity and evasion categories. To analyze the impact of class distribution on performance, we evaluated multiple data distribution strategies as follows:

- **Full Dataset:** We first trained and evaluated the models on the original dataset, thereby preserving the inherent class distribution defined in the shared task.
- **Rebalancing:** We applied class-balanced rebalancing by setting a target of 250 instances per class. Minority classes were oversampled with replacement, while majority classes were undersampled without replacement (seed=42), producing a balanced training set.

- **Random Undersampling:** To examine strict majority-class reduction, we applied random undersampling by limiting each class to 79 instances, the size of the smallest class in the original training set. For each label, 79 samples were randomly selected without replacement (seed=42), producing a balanced but reduced dataset.

- **LLM-based Paraphrastic Refinement:** In this setting, we rebalanced the dataset using paraphrased versions generated by an LLM. We paraphrased only the answer column of selected minority-class instances while preserving the original questions and labels. Majority classes, such as *Explicit*, were reduced from 1052 to 600 instances and *Dodging* from 706 to 500 to better match smaller classes. Minority classes-*Declining to answer* from 145 to 222, *Claims ignorance* from 119 to 214, *Clarification* from 92 to 155, and *Partial/half-answer* from 79 to 155-were expanded through controlled paraphrasing. Classes with moderate frequencies (*Implicit*: 488, *General*: 386, *Deflection*: 381) remained unchanged. The final dataset is shuffled before training. This controlled redistribution improves class balance while preserving semantic intent and label consistency. We used GPT-5.1 (Singh et al., 2025) to paraphrase answers. To ensure quality, a human-in-the-loop verification step was conducted to validate the outputs before inclusion in the final dataset. The prompt used for LLM-based paraphrasing is provided in Appendix F.2, and examples of paraphrased data points are shown in Appendix D. The augmented dataset is also publicly available on our GitHub repository.

4.3 Decoder-only Models

We evaluated four decoder-only LLMs: Qwen3-14B (Yang et al., 2025), Phi-4 (Abdin et al., 2024), Gemma-2 9B (Team et al., 2024), and Mistral 7B (Chaplot, 2023) across multiple training paradigms to analyze prompt-based learning, supervised fine-tuning, and the effects of data balancing strategies on both subtasks. We selected these models to represent a diverse set of widely used decoder-only LLMs for a fair comparison across learning settings. All fine-tuning experiments were conducted on base models, while instruction-tuned (-it) variants were used for zero-shot and few-shot (ICL) evaluations.

- **Zero-shot Learning:** We first evaluated the models in a zero-shot setting using an instruction-style prompt that included a question-answer pair and a predefined evasion taxonomy with brief label definitions. The full prompt is provided in Appendix F.3.

- **Few-shot Learning:** We further evaluated the models in a few-shot setting using an instruction-style prompt with labeled examples. Each input included a Question-Answer pair, a fixed evasion taxonomy, detailed definitions, and two illustrative examples per label. The full prompt is available in Appendix F.4.

- **Fine-tuning:** We further evaluated the models via supervised fine-tuning on the original dataset and three class-rebalancing strategies. We employed the UnSloth framework with Low-Rank Adaptation (LoRA) (Hu et al., 2022) to enable memory-efficient and scalable adaptation of LLMs. The training prompt is available in Appendix F.1.

The official submission achieved a Macro F1 score of 74.77% (ranked 21st) on Subtask-1 using Phi-4 trained on the original dataset for 3 epochs. For Subtask-2, our submission achieved 44.17% (ranked 22nd) using Phi-4 trained on the oversampled dataset for 5 epochs. Post-submission experiments revealed an improved model for Subtask-2, which is reported in this paper. We also conducted preliminary experiments training directly on the Subtask-1 objective; however, these yielded inferior performance compared to the hierarchical approach, consistent with prior findings Thomas et al. (2024).

Appendix A details the hyperparameter configurations and implementation settings for all decoder-only model experiments.

5 Results and Discussion

Table 2 presents the performance of different methods, evaluated using macro F1 score on the official evaluation-stage test data. The results offer a comparative analysis across the approaches, highlighting their relative strengths and potential limitations.

Phi-4 Achieves Consistently Strong Performance. Across all configurations, Phi-4 consistently achieved the strongest performance. On the

Model	Epochs	Subtask-1	Subtask-2
LLM-Based Paraphrastic Refinement Dataset			
Qwen3 14B	3	59.98	32.46
Qwen3 14B	4	61.22	36.00
Phi-4	3	**74.77**	48.60
Phi-4	4	73.85	**51.55**
Gemma-2 9b	3	70.41	48.97
Gemma-2 9b	4	72.65	47.22
Mistral 7b	3	70.23	47.03
Mistral 7b	4	69.54	46.18
Full original dataset			
Qwen3 14B	3	65.48	38.89
Phi-4	3	**74.54**	**48.62**
Gemma-2 9b	3	69.15	42.98
Mistral 7b	3	68.59	45.35
Rebalanced Dataset			
Qwen3 14B	3	58.98	29.85
Qwen3 14B	5	62.40	34.01
Phi-4	3	64.41	39.80
Phi-4	5	**69.20**	**44.17**
Gemma-2 9b	3	66.16	41.99
Gemma-2 9b	5	63.97	39.58
Mistral 7b	3	66.07	41.86
Mistral 7b	5	68.37	40.31
Undersampled Dataset			
Qwen3 14B	3	49.52	22.92
Qwen3 14B	10	59.12	36.62
Phi-4	3	**62.63**	36.22
Phi-4	10	59.12	36.18
Gemma-2 9b	3	36.90	17.99
Gemma-2 9b	10	54.07	34.38
Mistral 7b	3	50.54	35.78
Mistral 7b	10	55.76	**37.12**
Zero-shot			
Qwen3-14B-it	–	17.59	06.54
Phi-4	–	**70.80**	**39.88**
Gemma-2 9b-it	–	20.25	07.19
Mistral 7b-it	–	17.59	06.54
Few-shot			
Qwen3-14B	–	17.59	06.54
Phi-4	–	**64.91**	**45.64**
Gemma-2 9b-it	–	18.66	06.52
Mistral 7b-it	–	56.23	31.31

Table 2: Model performance comparison on Subtask-1 (Clarity Classification) and Subtask-2 (Evasion Classification), using Macro F1, reported in %.

LLM-augmented dataset, it reached 74.77% on Subtask-1 (3 epochs) and 51.55% on Subtask-2 (4 epochs). This advantage remained on the full original dataset (74.54% and 48.62%), as well as in zero-shot and few-shot settings, highlighting its strong generalization capabilities. Overall, Phi-4's architecture and pre-training appear better aligned with the target tasks than those of the other evaluated LLMs.

LLM-Based Paraphrastic Refinement Shows Mixed Impact on Performance. LLM-based refinement contributed to overall performance gains across multiple models when compared to training on the full original dataset, with all comparisons made under identical 3-epoch settings. Phi-4 at showed improvement on Subtask-1 (Δ = +0.23%) while maintaining nearly identical performance on Subtask-2 (Δ = -0.02%). Phi-4 achieved almost the same performance on both subtasks, with $\Delta = +0.23\%$ on Subtask-1 and $\Delta = -0.02\%$ on Subtask-2. Gemma-2 9B benefited more clearly from refinement, improving by Δ = +1.26% on Subtask-1 and Δ = +5.99% on Subtask-2. Mistral 7B achieved consistent improvements across both subtasks (Δ = +1.64% and Δ = +1.68%). For Qwen3-14B, refinement led to a decrease of Δ = -5.50% on Subtask-1 and Δ = -6.43% on Subtask-2. Overall, these findings demonstrate that controlled paraphrase refinement can meaningfully improve model performance. Qwen3-14B gained performance when trained for 4 epochs, with $\Delta = +1.24\%$ on Subtask-1 and $\Delta = +3.54\%$ on Subtask-2, while Gemma-2 9B gained $\Delta = +2.24\%$ on Subtask-1 but slightly decreased by $\Delta = -1.75\%$ on Subtask-2. While LLM-based paraphrasing improved class balance, it may also introduce subtle distribution shifts or stylistic biases that affect model learning. In some cases, these effects offset the gains from balancing minority classes, especially when the original dataset already captures the task distribution well. As a result, model performance appears to depend more on data quality and alignment than on refinement alone.

Rebalancing Fails to Improve Performance, While Undersampling Severely Degrades It. Training with a rebalanced dataset for 3 epochs did not improve performance over the full dataset baseline: Qwen3-14B declines by Δ= -6.50% on Subtask-1 and Δ = -9.04% on Subtask-2. Phi-4 drops by Δ = -10.13% and Δ = -8.82%, respectively. Gemma-2 9B decreases by Δ = -2.99% on Subtask-1 and Δ = -0.99% on Subtask-2. Mistral 7B shows reductions of Δ = -2.52% and Δ = -3.49% across Subtasks 1 and 2. These results suggest that redundancy limits generalization. Training with undersampling for 3 epochs degraded performance compared to the full-dataset setting. Qwen3-14B declined by Δ = -15.96% on Subtask-1 and Δ = -15.97% on Subtask-2. Phi-4 dropped by Δ = -11.91% and Δ = -12.40%. Gemma-2 9b experienced the largest degradation,

with Δ = -32.25% on Subtask-1 and Δ = -24.99% on Subtask-2. Overall, reduced data volume outweighed any benefits from class rebalancing, rendering undersampling ineffective across both subtasks. Although extended training in both rebalanced and undersampled settings partially mitigated these declines for some models, none recovered to baseline levels.

Zero-Shot and Few-Shot Inference Fall Short of Fine-Tuned Performance. Without task-specific fine-tuning, most models struggled to capture the task structure. Qwen3-14B-it, Gemma-2 9B-it produced similarly low scores in both zero-shot and few-shot settings, trailing their fine-tuned baselines by over Δ = -47% on Subtask-1 and Δ = -38% on Subtask-2. Few-shot prompting provided negligible gains (less than Δ = +1.07%). Mistral 7B-it achieved 17.59% and 6.54% in zero-shot, which increased significantly to 56.23% and 31.31% in few-shot (Δ = +38.64%, +24.77%), yet still lagged behind fine-tuned performance. Phi-4 achieved 70.80% and 39.88% in zero-shot, indicating stronger intrinsic alignment with the label schema. Its few-shot results improved Subtask-2 by Δ = +5.76% but decreased Subtask-1 by Δ = -5.89%, showing inconsistent benefits from added examples. It observes that the maximum output token limit was initially set to 16; to investigate whether this constraint contributed to the poor ICL performance, we increased it to 128, but observed no change in the results. Overall, these findings highlight that supervised fine-tuning is crucial for reliable performance, and ICL alone is insufficient without strong prior task alignment. The details of the evaluation metrics and sample predictions for both subtasks are provided in Appendices B and E, respectively. Appendix C illustrates the error analysis of the best-performed model.

6 Conclusion

In this work, we explored response clarity and evasion detection using multiple decoder-only LLMs under zero-shot, few-shot, and LoRA-based fine-tuning settings. Our results show that task-specific fine-tuning is essential for reliable performance, as most models struggled under in-context learning alone. From a data-centric perspective, simple resampling proved ineffective: rebalancing introduced redundancy, whereas undersampling significantly degraded performance. In contrast, controlled LLM-based paraphrasing led to performance improvements. Overall, our findings highlight that effective dataset design and class-distribution strategies play a more critical role than model scale in clarity and evasion classification.

Limitations

Despite promising findings, this study has several limitations. First, the dataset size remains relatively small, particularly for low-frequency evasion categories, which may limit generalization. Second, although paraphrastic refinement improves class balance, LLM-generated rewrites may introduce subtle stylistic biases or distribution shifts that affect downstream learning. Third, all experiments rely on decoder-only models; encoder-based or hybrid architectures were not explored. Future work should explore larger-scale datasets, cross-domain generalization, multilingual evaluation, and more advanced refinement or augmentation techniques to further improve robustness.

References

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

Peter Bull. 2003. *The microanalysis of political communication: Claptrap and ambiguity*. Routledge.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political psychology*, pages 651–666.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Jim T Dillon. 2025. *The practice of questioning*. Routledge.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors.

2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Codruta Liliana Girlea. 2017. *Deception detection in dialogues*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Sandra Harris. 1991. Evasive action: How politicians respond to questions in political interviews. *Broadcast talk*, 7699.

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

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141.

Rajkumar Pujari and Dan Goldwasser. 2021. Understanding politics via contextualized discourse processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Daniela Trotta and Sara Tonelli. 2021. Are gestures worth a thousand words? an analysis of interviews in the political domain. In *Proceedings of the 1st workshop on multimodal semantic representations (MMSR)*, pages 11–20.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Appendix

A Hyperparameter Configurations

For the LoRA fine-tuning approach, we utilized the Unsloth framework with 4-bit quantization to ensure memory-efficient training and a maximum sequence length of 4096 tokens. LoRA adapters were configured with $r = 32$ and $\alpha = 32$ to target the projection layers (query, key, value, output, gate, up, and down projections). LoRA dropout was set to 0, and gradient checkpointing was enabled using the optimized Unsloth mode. Models were trained with a learning rate of 2×10^{-4}, using the AdamW 8-bit optimizer. A linear learning rate scheduler with 5 warmup steps and weight decay of 0.001 was applied. Models were instructed to output the predicted label strictly within the <LABEL> </LABEL> tags. For inference, we generated model outputs using greedy decoding to ensure deterministic label predictions. During generation, we set max_new_tokens=16, disabled sampling (do_sample=False) with a single beam (num_beams=1), and left sampling-related parameters (temperature, top_p, top_k) unset.

B Evaluation Metric

System performance for both subtasks was evaluated using the macro F1-score as the official metric. Macro F1 computes the F1-score for each class independently and then averages them equally across all classes, as formulated in Eq. (4).

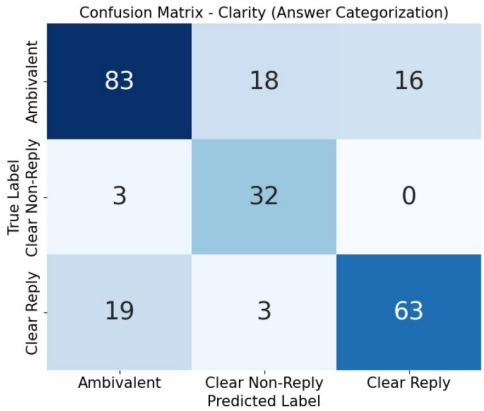
$$F1_{\text{macro}} = \frac{1}{L} \sum_{l=1}^{L} \frac{2 \cdot P_l \cdot R_l}{P_l + R_l} \tag{4}$$

where P_l and R_l denote precision and recall for class l, and L is the number of classes (3 for Subtask-1 and 9 for Subtask-2). This metric ensures balanced evaluation by giving equal weight to both frequent and rare classes.

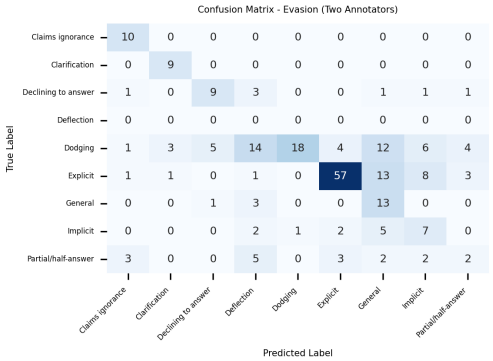
C Error Analysis

In Subtask-1 (Figure C.1a), misclassifications were concentrated around the Ambivalent category. Of the 117 true Ambivalent instances, 18 were predicted as Clear Non-Reply and 16 as Clear Reply. This bidirectional confusion aligns directly with findings reported by Thomas et al. (2024), where inter-annotator agreement between Ambivalent and the other two categories was substantially lower ($\kappa = 0.65$ with Clear Reply and $\kappa = 0.71$ with Clear Non-Reply). The confusion mainly comes from ambivalent answers, which can often be seen as either somewhat helpful or fully informative. In contrast, clear non-replies are the easiest to distinguish, with no cases being mistaken for clear replies.

In Subtask-2 (Figure C.1b), most misclassified *Dodging* instances were predicted as *General* or *Deflection*, consistent with Thomas et al. (2024), who report low agreement between these categories ($\kappa = 0.57$ and $\kappa = 0.62$). Their conceptual overlap and shared surface cues (e.g., vague phrasing, topic shifts) likely cause the model to rely on lexical patterns rather than intent. *Explicit* responses were also confused with *General* and *Implicit*, mirroring low agreement ($\kappa = 0.58$). Similarly, *Declining to answer* was mistaken for *Deflection* ($\kappa = 0.77$), as both may involve refusal-like language (e.g., I cant comment) or shifting responsibility (thats for others to decide). Finally, the low-frequency *Partial/half-answer* category was often absorbed into more frequent labels, reflecting limited training data and its hybrid nature.



(a) Clarity Prediction (Subtask 1)



(b) Evasion Prediction (Subtask 2)

Figure C.1: Confusion matrices for the best model (Phi-4 with distribution-aware refinement). (a) Clarity-level classification. (b) Fine-grained evasion-type classification.

D Paraphrased Dataset Samples

Table D.1 presents representative examples from our LLM-based paraphrastic refinement stage, where only the answer text in each instance is paraphrased while the associated question, clarity, and evasion labels are kept unchanged.

Original Answer	**Paraphrased Answer**	**Clarity Label**	**Evasion Label**
I'll let other people analyze that. But what I will emphasize is that one of the nice things about being in the sixth year of your Presidency is that you've seen a lot of ups and downs, and you've gotten more than your fair share of attention. And I've had the limelight, and I'vethere have been times where the requests for my appearances were endless. There have been times when, politically, we were down. And it all kind of evens out, which is why what's most important, I think, is keeping your eye on the ball, and that is, are you actually getting some good done? Scott Horsley [National Public Radio], last question.	I'll leave that for others to interpret. What I want to stress, though, is that one of the benefits of being in the sixth year of a presidency is having experienced plenty of highs and lows, along with more than enough attention. Ive had my time in the spotlight, and there were moments when requests for appearances never seemed to end. There were also times when we were politically struggling. In the end, it all balances out. Thats why the most important thing is staying focused on what really matterswhether youre actually accomplishing something worthwhile. Scott Horsley from NPR, final question.	Clear Non-Reply	Declining to answer
You know, on the first question, I can't answer that. I'm not going to give you intelligence data, number one. Number two, we would respond. We would respond if he uses it. The nature of the response would depend on the use.Josh [Josh Wingrove] of Bloomberg.	On your first question, I cant address that because I wont share intelligence information. What I can say is that we would respond if he used it, and the nature of that response would depend on how he used it.	Ambivalent	Partial/half-answer
But read the reports. China issued a statement that it was an airborne disease. I heard it was an airborne disease. I assumed it early on. The fact is that there must be calm. You don't want me jumping up and down, screaming: There's going to be great death. There is something going on that is really causing very serious problems for the country. If Bob Woodward thought what I said was bad, he should have immediately, right after I said it, gone to the authorities so they could prepare and notify them.	But look at the reports. China issued a statement saying it was an airborne disease. I heard early on that it was airborne, and I assumed that from the beginning. The reality is that there needs to be a sense of calm. You dont want me jumping up and down, shouting that theres going to be massive death and panicthat would have caused extremely serious problems for the country. If Bob Woodward believed what I said was so alarming, then he should have gone straight to the authorities at the time and informed them so they could prepare and take action.	Clear Non-Reply	Claims ignorance

Table D.1: Example format for original and paraphrased answers with clarity and evasion labels

E Prediction Examples

Tables E.1 and E.2 present sample predictions for the two subtasks. In Table E.1, example question-answer pairs are shown alongside their predicted and actual clarity labels, where predictions are obtained using the Phi-4 model trained for 3 epochs on the augmented dataset. In Table E.2, sample question-answer pairs are presented with their corresponding predicted and actual evasion labels, generated using the Phi-4 model trained for 4 epochs on the augmented dataset.

Question–Answer Pair	**Predicted Clarity Label**	**Actual Clarity Label**
Sample1: Q: Do you think that is a good thing? **A:** THE PRESIDENT. I think it's a good thing. I was very lucky. When I went to the Senate, I got three major committees. I was put on the Appropriations Committee, on the Interstate Commerce Committee	Clear Reply	Clear Reply
Sample2: Q: Does the Prime Minister feel that that would have a major impact on the peace talks? **A:** The President. Nice try. The Prime Minister. I don't deal with domestic political American or personal domestic American problems.	Clear Non-Reply	Ambivalent
Sample3: Q: Is that in public? **A:** President Trump. Okay? Do you have a question here?	Ambivalent	Ambivalent

Table E.1: Sample predictions with actual and predicted labels for clarity classification.

Question–Answer Pair	**Predicted Evasion Label**	**Actual Evasion Label**
Sample1: Q: Do you think that is a good thing? **A:** THE PRESIDENT. I think its a good thing. I was very lucky. When I went to the Senate, I got three major committees. I was put on the Appropriations Committee, on the Interstate Commerce Committee	Explicit	Explicit
Sample2: Q: Does the Prime Minister feel that that would have a major impact on the peace talks? **A:** The President. Nice try. The Prime Minister. I dont deal with domestic political American or personal domestic American problems.	Declining to answer	Dodging
Sample3: Q: Is that in public? **A:** President Trump. Okay? Do you have a question here?	Dodging	Dodging

Table E.2: Sample predictions with actual and predicted labels for evasion-type classification.

F Prompt Design

F.1 Prompt used for training

Table F.1 presents the prompt used during supervised fine-tuning for evasion-level classification. The prompt guides the model to evaluate how directly a given answer addresses its corresponding question and to generate the appropriate evasion label accordingly. To ensure consistent, structured outputs, the model is instructed to produce only the selected label, enclosed within predefined <LABEL> tags, without any additional explanation or reasoning.

Prompt used for zero-shot learning for evasion-level classification

You are classifying the type of answer given to a question.
Question: question
Answer: answer
First, decide how directly the answer addresses the question. Then, output ONLY the evasion label between <LABEL> </LABEL>.

Table F.1: Prompt used for training for evasion-level classification

F.2 Prompt Design for LLM-Based Text Paraphrasing

Table F.2 presents the structured prompt used for LLM-based paraphrasing. The prompt instructs the model to rewrite the given answer while strictly preserving its original meaning, intent, and approximate length. It explicitly prohibits adding new information, removing important details, altering factual content or stance, or significantly changing the length of the response. The model must output only the paraphrased answer, without any explanations or additional text.

Prompt Design for LLM-Based Text Paraphrasing

Your task is to paraphrase the answer while preserving:
-The original meaning and intent
-The overall length (do not significantly shorten or expand it; keep it approximately the same length)
-Do NOT add new information
-Do NOT remove important details
-Do NOT change the stance or factual content
-Do NOT make it substantially shorter or longer
Output only the paraphrased version of the answer. Do not include explanations, comments, or any additional text.
Answer: {answer}

Table F.2: Paraphraing prompt

F.3 Prompts used for Zero-shot training

Table F.3 illustrates the prompt design used for zero-shot learning in evasion-level classification. The prompt instructs the model to evaluate an answer strictly with respect to its corresponding question and to assign exactly one label from a predefined evasion taxonomy. The taxonomy comprises nine fine-grained evasion categories, each accompanied by a concise definition to ensure consistent interpretation. The prompt also enforces strict output formatting rules, requiring the model to produce only the selected label wrapped within predefined tags, without any explanation or additional text.

Prompt used for zero-shot learning for evasion-level classification

You are classifying the type of answer given to a question in a political interview.
Question: {question}
Answer: {answer}
Your task:
1) Judge the Answer ONLY with respect to the given Question.
2) Choose EXACTLY ONE evasion label from the taxonomy below.
Evasion label taxonomy (choose one):
- **Explicit**: The information requested is explicitly stated (in the requested form).
- **Implicit**: The information requested is given, but not explicitly stated (not in the expected form).
- **General**: The information provided is too general / lacks the requested specificity.
- **Partial/half-answer**: Provides only a specific component of the requested information.
- **Dodging**: Ignores the question altogether (does not acknowledge it, shifts to another topic).
- **Deflection**: Acknowledges the question but shifts focus and makes a different point than what is asked.
- **Declining to answer**: Acknowledges the question but refuses to answer at the moment (directly or indirectly).
- **Claims ignorance**: Claims/admits they do not know the answer.
- **Clarification**: Asks for clarification instead of providing the requested information.
Output format rules:
- Output ONLY the chosen label wrapped exactly like this: <LABEL>...</LABEL>
- Do NOT output explanations, reasoning, or any other text.
Now classify the Answer relative to the Question and respond with only:
<LABEL>YOUR_LABEL</LABEL>

Table F.3: Prompt used for zero-shot learning for evasion-level classification

F.4 Prompts used for Few-shot training

Table F.4 presents the prompt used for few-shot evasion-level classification with taxonomy description. In this version, each evasion category is accompanied by a concise definition and two illustrative questionanswer examples. These in-context examples are designed to guide the model toward a clearer understanding of the distinctions between closely related evasion strategies.

The prompt explicitly instructs the model to evaluate the answer strictly with respect to the given question and to select exactly one label from the nine-category taxonomy. Strict output formatting constraints are imposed, requiring the model to produce only the selected label wrapped within predefined tags, without any additional explanation or reasoning.

Prompt with taxonomy and examples for evasion-level classification

You are classifying the type of answer given to a question in a political interview.
Your task:
1) Judge the Answer ONLY with respect to the given Question.
2) Choose EXACTLY ONE evasion label from the taxonomy below.
TAXONOMY (definitions + 2 examples each)
1) Explicit
Definition: The information requested is explicitly stated (in the requested form).
Example 1
Question: What is the unemployment rate right now?
Answer: It is 6.1 percent in the latest labor report.
<LABEL>Explicit</LABEL>
Example 2
Question: Did you vote for the bill?
Answer: Yes, I voted for it.
<LABEL>Explicit</LABEL>
2) Implicit
Definition: The information requested is given, but not explicitly stated (not in the expected form).
Example 1
Question: Did you raise the corporate tax rate?
Answer: Corporations contribute more today than before our reform.
<LABEL>Implicit</LABEL>
Example 2
Question: Are you against the policy?
Answer: Ive been critical of it from the beginning.
<LABEL>Implicit</LABEL>
3) General
Definition: The information provided is too general / lacks the requested specificity.
Example 1
Question: How many new homes will be built under this plan?
Answer: Were going to build a lot of homes and make housing more affordable.
<LABEL>General</LABEL>
Example 2
Question: Which companies received the contracts?
Answer: A number of qualified firms were involved through a proper process.
<LABEL>General</LABEL>
4) Partial/half-answer
Definition: Provides only a specific component of the requested information.
Example 1

Question: Which two cities will receive the new hospitals?
Answer: One will be built in Chittagong.
<LABEL>Partial/half-answer</LABEL>
Example 2
Question: What are the three main causes of the delay?
Answer: One cause was a shortage of materials.
<LABEL>Partial/half-answer</LABEL>
5) Dodging
Definition: Ignores the question altogether (does not acknowledge it, shifts to another topic).
Example 1
Question: Did you meet with the lobbyist last week?
Answer: Our administration is focused on delivering results for the people.
<LABEL>Dodging</LABEL>
Example 2
Question: Did you use government funds for personal travel?
Answer: Let me tell you about the strong ethics reforms weve introduced.
<LABEL>Dodging</LABEL>
6) Deflection
Definition: Acknowledges the question but shifts focus and makes a different point than what is asked.
Example 1
Question: Why did you cut the education budget?
Answer: I understand the concern, but what matters is that student outcomes are improving.
<LABEL>Deflection</LABEL>
Example 2
Question: Did your office ignore the warning?
Answer: I hear what you're sayingthis is really about fixing the system so it never happens again.
<LABEL>Deflection</LABEL>
7) Declining to answer
Definition: Acknowledges the question but refuses to answer at the moment (directly or indirectly).
Example 1
Question: Will you resign if the investigation finds wrongdoing?
Answer: Im not going to discuss hypotheticals while the investigation is ongoing.
<LABEL>Declining to answer</LABEL>
Example 2
Question: Did you speak with the prosecutor?
Answer: I cant comment on that at this time.
<LABEL>Declining to answer</LABEL>
8) Claims ignorance
Definition: Claims/admits they do not know the answer.
Example 1
Question: What was the exact cost of the project?
Answer: I dont know the exact figure off the top of my head.
<LABEL>Claims ignorance</LABEL>
Example 2
Question: How many people were affected in total?
Answer: Im not sure of the exact number right now.
<LABEL>Claims ignorance</LABEL>
9) Clarification

Definition: Asks for clarification instead of providing the requested information.
Example 1
Question: When did you first learn about the allegations?
Answer: Which allegations are you referring to, and from which report?
<LABEL>Clarification</LABEL>
Example 2
Question: Did you approve the plan?
Answer: Which plan do you meanthe initial proposal or the revised version?
<LABEL>Clarification</LABEL>
Now classify the following:
Question: {question} Answer: {answer}
Output format rules:
- Output ONLY the chosen label wrapped exactly like this: <LABEL>...</LABEL>
- Do NOT output explanations, reasoning, or any other text.
Now classify the Answer relative to the Question and respond with only:
<LABEL>YOUR_LABEL</LABEL>

Table F.4: Prompt with taxonomy and in-context examples for evasion-level classification