SyntaxMind at SemEval-2026 Task 6: Exploring Transformers and LLMs for Unmasking Political Question Evasions

Md. Shihab Uddin Riad
Dept. of Computer Science & Engineering
International Islamic University Chittagong
shihab.riadn@gmail.com

Abstract

This paper describes our approach to Subtask 1: Clarity-level Classification in SemEval-2026 Task 6. The task focuses on determining the clarity of political responses with respect to their corresponding questions. To enhance model performance, we introduced a direct answer generation strategy as an additional input feature and applied Task-Adaptive Pre-Training (TAPT) to enhance encoder-only Transformer models with the task domain. We further explored both cross-entropy and focal loss to address potential class imbalance. Experimental results show that TAPT enhanced encoder models, particularly DeBERTa-V3-base, achieved the strongest performance, while generative small language models fine-tuned via parameter efficient methods exhibited comparatively lower results. Our system obtained a macro-F1 score of 0.72 on the official evaluation set, ranking 24th out of 40 teams.

1 Introduction

Any form of question evasion can be defined as a strategic approach to avoid providing direct answers to questions. In this regard, politicians have a long-standing tradition of avoiding direct responses to questions (Bull and Mayer, 1993; Íñigo-Mora and Deligiorgi, 2007). To address response ambiguity in political questioning, SemEval-2026 Task 6: CLARITY - Unmasking Political Question Evasions(Thomas et al., 2026) featured a novel, two-level taxonomy approach "I Never Said That": A dataset, taxonomy and baselines on response clarity classification (Thomas et al., 2024) paper.

We participated in Subtask 1: Clarity-level Classification. This subtask involves a multiclass text classification problem aimed at determining the clarity level of a given answer with respect to its corresponding question. The dataset paper primarily experimented with prompting and fine-tuning various large language models. While their work focused on large neural architectures, our study emphasizes the effectiveness of more resource efficient small language models less than or around 1B parameters. To extend their work, we introduced the following contributions:

- We expanded the dataset features by generating concise answers to questions from the given interviews using small language models.
- Encoder-only Transformer models were explored with Task-Adaptive Pre-Training (TAPT) to improve task-specific performance. Additionally, several generative small language models were fine-tuned to evaluate their effectiveness on the classification task.

Our key finding is that the TAPT enhanced models outperformed the large language model results reported in the dataset paper. However, the performance of standalone small language models was not competitive with the large language models evaluated in the original study.

2 Related Work

Research in question answering (QA) and related response interpretation has been an active area of natural language processing. Early QA systems focused on rule-based information retrieval and fact extraction techniques. In parallel, evaluation initiatives such as the TREC QA established foundational benchmarks for answer retrieval from text corpora (Voorhees and Tice, 2000).

The introduction of large pre-trained Transformer models such as BERT significantly advanced QA performance by contextualizing language representations through self-attention mechanisms (Devlin et al., 2018). Benchmarks such as SQuAD 2.0 (Rajpurkar et al., 2018) extended the task beyond extractive QA by incorporating

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3377–3381

unanswerable questions, thereby encouraging research into response quality, ambiguity, and answer informativeness. More recently, large language model approaches have been explored for generative QA and instruction-following tasks, with methods such as prompt-based fine-tuning and parameter-efficient adaptation (e.g., LoRA and PEFT) enabling effective deployment under limited computational resources (Dettmers et al., 2023).

3 System Overview

The objective of the system is to determine the clarity level of a given answer with respect to its corresponding question. The proposed system consists of three main components: (i) Direct Answer Extraction (ii) Task-Adaptive Pre-Training (TAPT) and (iii) supervised fine-tuning for multiclass classification.

3.1 Direct Answer Extraction

To enrich the input representation, an additional feature was generated by producing a concise, direct answer to each question from the corresponding interview context. This was accomplished using small language models with fewer than 2B parameters. Given an interview transcript containing a question–answer pair, then model was prompted to extract or generate a short, direct response to the question based on the interview content. The generated concise answer was then incorporated as an auxiliary feature alongside the original question–answer pair.

The motivation behind this step is that explicitly modeling a "direct answer" representation may help the classifier better capture discrepancies between the original response and a clear, focused reply. This additional feature aims to reduce ambiguity and improve clarity-level discrimination. The prompt used for direct answer extraction is shown in App. A.1.

3.2 Task-Adaptive Pre-Training (TAPT)

Recent studies (Gururangan et al., 2020) indicate that Task-Adaptive Pre-Training is less expensive to run than (Domain-Adaptive Pretraining) DAPT. To enhance task specific performance, encoder-only Transformer models underwent Task-Adaptive Pre-Training before supervised fine-tuning.

TAPT was performed using unlabeled task related data drawn from the same distribution as the shared task dataset, including a small amount of additional in domain data[1]. During this stage, the models continued pretraining using a masked language modeling objective to better adapt to the linguistic characteristics and discourse patterns of political interviews. By aligning the model parameters more closely with the task domain before fine-tuning, TAPT enhances contextual understanding and facilitates improved downstream classification performance.

3.3 Supervised fine-tuning

Following Task-Adaptive Pre-Training (TAPT), supervised fine-tuning was conducted on the labeled clarity level classification dataset. For encoder-only architectures, pre-trained Transformers (Wolf et al., 2020) models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DeBERTaV3 (He et al., 2021), ModernBERT (Warner et al., 2024) and ALBERT(Lan et al., 2019) were fine-tuned by adding a task-specific classification head on top of the representation. Both cross-entropy and focal loss (Lin et al., 2018) were investigated for multiclass classification, with focal loss considered to mitigate potential class imbalance.

In addition, several small instruction tuned generative language models were adapted using Parameter-Efficient Fine-Tuning (PEFT) techniques (Mangrulkar et al., 2022). Specifically, QLoRA (Dettmers et al., 2023) was employed, which combines Low-Rank Adaptation (LoRA) with 4-bit quantization through the bitsandbytes library. Under this framework, the base model weights remain frozen while only lightweight low-rank adapter parameters are updated, enabling memory efficient fine-tuning without substantial degradation in performance. This approach allows effective adaptation of language models under constrained computational resources. The prompt used for LoRA fine-tuning is shown in App. A.2.

4 Experimental Setup

Table 1 illustrates the hyperparameter setting of our trained models. We utilized the Kaggle platform for experimental purposes and implemented our model using Transformers Trainer and PyTorch library (Paszke et al., 2019).

[1] https://huggingface.co/datasets/shihab-0x0/presidents-news-conferences

Parameters	Value
Batch size	16
Epochs	5
Learning rate	2×10^{-5}
Loss function	CE Loss | Focal Loss
Optimizer	AdamW
Max sequence length	512
Warmup ratio	10%
Weight decay	0.01
Focal Loss Gamma	2.0
Focal Loss Alpha	0.25

Table 1: Hyperparameter values. (Abbreviations: CE = Cross-Entropy)

5 Results

Table 2 presents the official rankings of participating teams in SemEval-2026 Task 6 (Subtask 1). The proposed system, achieved an F1-macro score of 0.72, securing 24th position among 40 teams.

Teams	Rank	Score (F1-Macro)
TeleAI	1	0.89
AsymVerify	2	0.85
argha	20	0.75
SyntaxMind	**24**	**0.72**
uir_cis	35	0.61
lakksh	40	0.31

Table 2: Performance ranking of participating teams in SemEval-2026 Task 6 (Subtask 1) on evaluation dataset

Model (TAPT)	CE Loss	Focal Loss
BERT-base	0.53	0.50
RoBERTa-base	**0.54**	0.53
ModernBERT-base	0.45	0.46
DeBERTa-V3-base	0.53	**0.61**
ALBERT-V2-base	0.32	0.41

Table 3: Macro-F1 performance comparison of fine-tuned model results on test dataset

Table 3 reports the performance comparison of fine-tuned encoder-only Transformer models using cross-entropy (CE) and focal loss. Among the evaluated models, DeBERTa-V3-base achieved the best overall performance, reaching an F1-macro score of 0.61 with focal loss. Across most models, focal loss consistently improved performance compared to cross-entropy loss. This improvement suggests that addressing class imbalance plays a significant role in clarity-level classification.

Model	Score (F1-Macro)
LFM2.5-1.2B-Instruct	0.27
SmolLM2-360M-Instruct	0.33
Qwen2.5-0.5B-Instruct	0.26

Table 4: Performance comparison of fine-tuned small language model results on test dataset

In contrast, as shown in Table 4, fine-tuned small language models (LFM2.5-1.2B-Instruct (AI, 2025), SmolLM2-360M-Instruct (Allal et al., 2025) and Qwen2.5-0.5B-Instruct (Team, 2024)) exhibited substantially lower performance compared to encoder-only classifiers. The best performing SmolLM2-360M-Instruct model, achieved 0.33 macro-F1 which remains considerably below the encoder-based (TAPT) approaches. This indicates that, under limited parameter budgets and direct fine-tuning settings, small generative models may struggle with fine-grained clarity-level classification compared to specialized encoder architectures.

5.1 Error Analysis

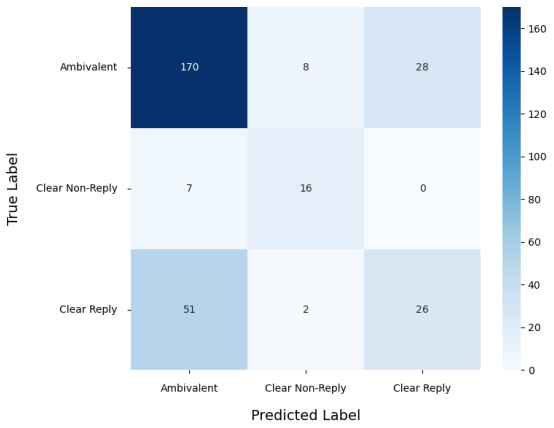

Figure 1: Confusion matrix of DeBERTa-V3-base with focal loss on the test dataset for SemEval-2026 Task 6 (Subtask 1)

The confusion matrix in Figure 1 reveals several systematic misclassification patterns across the three classes. The model performs strongest on the Ambivalent class, correctly predicting 170 instances out of 206.

For the Clear Non-Reply class, performance is comparatively weaker. While 16 instances are correctly identified, misclassification into Ambivalent (7 cases) suggests that the model struggles to distinguish between vague responses and explicit non-answers. The most significant challenge is observed in the Clear Reply class. A large portion

of true clear replies (51 cases) are incorrectly predicted as Ambivalent, indicating that the model tends to underestimate clarity and favors a more neutral or uncertain classification. This bias may arise from class imbalance or insufficient modeling of explicit answer patterns.

6 Conclusion

This paper presented a system for Clarity-level Classification in SemEval-2026 Task 6. The proposed approach implements direct answer generation as an additional input feature and leveraged Task-Adaptive Pre-Training (TAPT) to enhance domain alignment. Experimental results demonstrate that encoder-only Transformer models achieved the strongest performance, with focal loss further improving robustness under class imbalance. In future work, we intend to improve small language model performance so that they can achieve more competitive results while maintaining computational efficiency.

References

Liquid AI. 2025. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14(4):651.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

I. Íñigo-Mora and K. Deligiorgi. 2007. Evasion in political interviews: An analysis of televised interviews with Tony Blair. *Political Linguistics*, 23(3):78–90.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

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. *CoRR*, abs/1907.11692.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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. *Preprint*, arXiv:2409.13879.

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.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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.

A Prompting Details

A.1 Direct Answer Extraction Prompt

System Prompt

You are an information extraction assistant. Given a question and an interview transcript, extract the full contiguous portion of the transcript that constitutes the responder's answer.

Rules

- Do not select only a single sentence if more answer-related text follows
- Do not repeat the question
- Do not add attribution phrases
- Do not paraphrase, summarize, or explain
- Use only words that appear verbatim in the transcript
- Include indirect, evasive, qualifying, or extended responses
- If the answer spans multiple sentences, include all of them
- Return the extracted answer text

User Prompt

```
QUESTION: {question}
TRANSCRIPT: {interview_answer}
```

A.2 Prompt for LoRA fine-tuning

System Prompt

You are a classification assistant. Given an interview question and its answer, classify the answer into exactly one of the following categories:

- Clear Reply
- Clear Non-Reply
- Ambivalent

Respond with only the category name.

User Prompt

```
Question: {question}
Answer: {answer}
Interview Question: {interview_question}
Interview Answer: {interview_answer}
```