ChulaNLP at SemEval-2026 Task 6: A Hybrid BERT–LLM Framework for Political Response Clarity and Evasion Detection

Wisarut Peerachaidecho
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
wisarut.mos@gmail.com

Attapol T. Rutherford*
Department of Linguistics
Faculty of Arts
Chulalongkorn University
attapol.t@chula.ac.th

Abstract

SemEval-2026 Task 6 (CLARITY: Unmasking Political Interview) focuses on detecting equivocation and evasion techniques in political interviews. While encoder-only models and Large Language Models (LLMs) individually struggle with this task, we propose a hybrid BERT–LLM framework to leverage their complementary strengths: the discriminative efficiency of fine-tuned encoders and the sophisticated reasoning of LLMs. We benchmarked several long-context architectures—DeBERTa, RooseBERT, and BigBird—finding that a truncated DeBERTa-large provided the most reliable candidates for the LLM. By using DeBERTa's top-5 predicted labels as constrained options for LLM inference, we significantly improved evasion-level classification. This hybrid approach achieved competitive rankings in the shared task, placing 7th in Subtask 1 and 2nd in Subtask 2.

1 Introduction

In the contemporary media environment, communication often unfolds widely and rapidly over the online platform. Within the political media context, for example, politicians frequently refrain from providing direct answers during interviews, particularly when confronted with sensitive, controversial, or strategically disadvantageous questions. Rather than addressing the substance of a question, speakers may shift topics, provide overly general statements, or ignore the question altogether. Such strategies allow politicians to manage risk, protect their public image, and maintain strategic ambiguity. This behavior, known as equivocation or evasion, is a form of communication where responses are intentionally unclear. (Dillon, 2025)

To study this phenomenon, Thomas et al. (2026) introduced the Response Clarity Evaluation task. SemEval-2026 Task 6 builds on the Response Clarity Evaluation task introduced by this work, which investigates whether a politician directly answers a question and, if not, what type of evasion strategy is used. The task is framed as a hierarchical classification problem over question–answer pairs from political interviews. In SemEval-2026 Task 6, Subtask 1 (Clarity Level) requires systems to determine how clearly a response answers a question, classifying it as a clear answer, an ambivalent answer, or a clear non-response. Subtask 2 (Evasion Level) then asks systems to identify the specific evasion technique employed when the response is not fully clear, choosing from nine categories such as explicit answering, partial or half answering, implicit responses, dodging, deflection, clarification, declining to answer, or claiming ignorance. Together, the two subtasks evaluate both the overall clarity of a response and the particular strategies used to avoid directly addressing a question.

To address this task, we propose a hybrid BERT-LLM framework, which combine the encoder-only language models, which learn effectively from task-specific datasets, and large language models, which possess a superior general language understanding capability. Specifically, we investigate whether encoder-only architectures, when fine-tuned on the training data, can produce reliable and data-efficient predictions, and how these predictions can be used to guide a more general-purpose LLM. We therefore propose a hybrid framework in which the outputs of encoder-only models serve as structured guidance for an LLM, combining task-specific discriminative strength with large-scale generative knowledge.

We propose a system for SemEval-2026 Task 6 based on this framework and evaluate a wide of range of modern encoder-only language models and also (decoder-only) large language models. Our results show that DeBERTa-large trained on truncated question–answer pairs performs best

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1874–1881

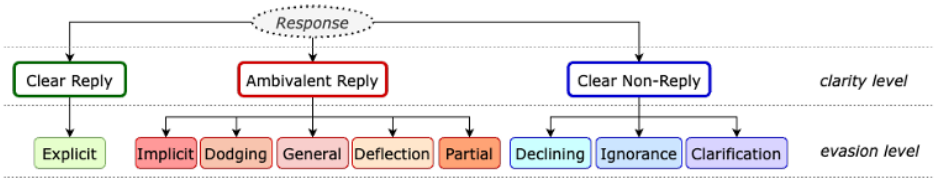

Figure 1: Taxonomy of response clarity classification proposed by Thomas et al. (2024)

among encoder-only models. Combining its top-5 predicted labels as constrained candidate options for Kimi-K2 significantly boosts the LLM's performance on the second subtask.

Our ChulaNLP team, registered under the Codabench username `moswisarut` [1] ranked 11th and 4th in the initial evaluation phase for Subtasks 1 and 2, respectively. In the final evaluation phase, our rankings improved to 7th place for Subtask 1 and 2nd place for Subtask 2.

2 Related Work

The task of response clarity evaluation was first formalized by Thomas et al. (2024), who examined how the interviewees articulate the answers to politically oriented questions. Their work framed the problem as determining whether a response is delivered with clarity or ambiguity and, when ambiguity is present, identifying the specific evasion technique used. This formulation positioned response clarity as a measurable linguistic property and highlighted the need for systematic, annotation-driven approaches to studying evasive communication in interviews.

To support this task, Thomas et al. (2024) introduced a two-level hierarchical typology designed to capture coarse and fine grain distinctions in response behavior as shown in Figure 1. At the top level, responses are categorized into three clarity classes. Clear replies represent direct and unambiguous answers that permit only a single reasonable interpretation. Ambivalent replies include responses that remain relevant to the question, but allow multiple interpretations due to vague or qualified phrasing. Clear non-replies, by contrast, consist of explicit refusals to provide information or engage with the question posed. These three categories are further expanded into nine subtypes of evasion techniques, each corresponding to a particular strategy of ambiguity or avoidance. This hierarchical structure enables a more nuanced characterization of how political figures navigate questions and manage the expectations of interviewers and audiences.

The dataset underlying this typology was created through a multi-step human annotation process. Because political interviews often contain compound questions, the original question–answer turns were first decomposed into singular question–answer pairs to ensure that each response could be evaluated against a single, well-defined query. These refined pairs were then assigned to trained annotators who labeled each instance according to the proposed taxonomy. This procedure ensured annotation consistency and facilitated a reliable mapping between linguistic responses and their corresponding clarity categories and evasion techniques.

BERT and other encoder-only language models are designed for discriminative tasks and can be fine-tuned efficiently using relatively small amounts of labeled data (Devlin et al., 2019). Because they are trained with task-specific supervision through classification objectives, they directly optimize for label prediction, making them data-efficient and stable during fine-tuning. In contrast, large language models are often times too cumbersome to fine-tune for one specific task without catastrophic forgetting. Standard BERT models are limited to 512 input tokens, which restricts their ability to model long documents where relevant information may span multiple paragraphs. Moreover, the quadratic complexity of full self-attention makes scaling to longer sequences computationally expensive. Zaheer et al. (2020) proposed BigBird, which supports sequences up to 4096 tokens using a block sparse attention mechanism that combines local, global, and random attention. BigBird has demonstrated strong performance on long-document tasks such as summarization and question answering. DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is an enhanced transformer-based encoder model that improves upon BERT by introducing disentangled attention and an improved mask decoder (He et al., 2021).

[1] Our codes are available at https://github.com/moswisarut/SemEval2026-Task6-moswisarut

Dore et al. (2025) introduced RooseBERT, a family of BERT-based models pretrained specifically on political discourse. The motivation is that political language contains domain-specific rhetorical and strategic patterns that are not well represented in general corpora. RooseBERT is pretrained on a large collection of political debate transcripts, making it particularly suitable for downstream tasks involving political text such as ours.

3 Our Approach

We framed political evasion classification as a multi-class problem by concatenating each QA pair and feeding it to DeBERTa, RooseBERT, or BigBird, with inputs truncated to model-specific token limits. After fine-tuning all encoders, we selected the best-performing model and used its top-3 and top-5 label predictions as constrained candidate sets for LLM inference. DeepSeek and Kimi-K2 then performed zero-shot or few-shot classification using either the full label set or these reduced candidate sets. Their outputs served as the final predictions for Subtask 2 and were further mapped to broader clarity categories for Subtask 1.

Writing mathematically, let $x = [q; a]$ denote the concatenated QA pair for each dialogue turn. For each encoder model $M \in \{\text{DeBERTa}, \text{RooseBERT}, \text{BigBird}\}$, we apply the corresponding tokenizer T_M to obtain a token sequence $s = T_M(x)$. We enforce a maximum input length L_M, where $L_{\text{BERT}} = 512$ and $L_{\text{BigBird}} = 4096$, by truncating s to $\tilde{s} = s_{1:L_M}$. Each model is fine-tuned with a multi-class classification head to optimize the cross-entropy objective

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log p(c \mid \tilde{s}, M),$$

where $C = 9$ is the number of evasion subcategories. From the highest-performing encoder model M^*, we compute the sorted probability vector $p^* = p(\cdot \mid \tilde{s}, M^*)$ and extract the top-k label sets $L_k = \text{TopK}(p^*, k)$ for $k \in \{3, 5\}$. These candidate sets L_k, along with the full label set $\mathcal{C}$, are used as constrained label spaces for large language models $G \in \{\text{DeepSeek}, \text{Kimi-K2}\}$ to perform zero-shot or few-shot inference. The LLM prediction $\hat{c}_2 = G(x, L_k)$ serves as the output for Subtask 2. Using the taxonomy mapping function $h : \mathcal{C} \to \mathcal{Z}$ from evasion subcategories to clarity levels, we compute the corresponding clarity prediction $\hat{z}_1 = h(\hat{c}_2)$ for Subtask 1.

4 Experimental Setup

In this SemEval-2026 shared task, all participants were provided with the QA pairs from the presidential interview dataset introduced by Thomas et al. (2024). The dataset was constructed by scraping official interview transcripts from the White House website, which are all in English, followed by manual annotation of evasion techniques. The training split consists of 3,448 singular QA pairs; however, only 3,390 of these are unique. Among the unique instances, 3,332 pairs were annotated by a single annotator, while 58 question or sub-question pairs were independently annotated by two annotators, resulting in 116 duplicated QA entries. These duplicates can either offer redundant signals—when annotators agree—or introduce noise when annotators disagree. To ensure training stability, we retained only duplicated instances where both annotators assigned the same evasion category and removed those with mismatched labels. This filtering step resulted in keeping 32 duplicated QA pairs and discarding 26, yielding a final training set of 3,364 high-quality QA instances. The development set contains 308 singular QA pairs.

A substantial portion of the dataset contains question–answer pairs whose concatenated text exceeds the 512-token input limit of DeBERTa when tokenized with the RoBERTa tokenizer. In the training data set, the sequence lengths reach up to 2,500 tokens and up to 1,750 tokens in the development set. Consequently, 1,261 training instances (36.57%) and 134 development instances (43.51%) surpass the 512-token threshold, indicating that long-context handling is necessary for this task.

4.1 Baseline Model

For our baseline systems, we fine-tune DeBERTa-base and DeBERTa-large on a filtered subset of the training data. Specifically, we retain only those QA pairs whose concatenated sequences do not exceed 512 tokens, following the stability-enhancing filtering strategy reported in Thomas et al. (2024). We report this preprocessing strategy as "Drop".

4.2 Experimental Model

We fine-tune the DeBERTa models on the full training set, truncating samples longer than 512 tokens instead of removing them, thereby maximizing data utilization. We report this preprocessing strategy as "Truncate". Second, we evaluate RooseBERT-base-cased-cont, a model pretrained on political-

Labels	Model	Method	F1 (S2)	F1 (S1)
All 9	DeBERTa-large	Fine-tuned	0.43	0.76
All 9	Kimi-K2	Few-shot	–	0.66
Top 5	Kimi-K2	Few-shot	**0.61**	**0.82**

Table 1: Model inference performance on Subtask 1 and Subtask 2 of the official test set. All macro F1-scores are from the Codabench evaluation leaderboard.

debate corpora, to test whether domain-specific pretraining offers advantages over general-purpose encoders. Third, to study the effects of longer contextual windows, we experiment with the BigBird-RoBERTa-base, which accommodates sequences up to 4096 tokens, enabling substantially broader context modeling compared to standard 512-token architectures. Since we allowed the token sequence length up to 4096 tokens, we report this preprocessing strategy as "Allow".

We also evaluate recently released LLMs—DeepSeek (DeepSeek-V3.2-Exp) and Kimi-K2 (kimi-k2-0905-preview) under zero-shot and few-shot settings using structured batch classification prompts.

Finally, we explore a hybrid inference approach in which LLM predictions are constrained by the top-3 and top-5 candidate labels generated by our best-performing encoder-only model, thereby combining discriminative model precision with the generalization capabilities of LLMs. Due to budget constraints, we only choose the best performing LLM to experiment our hybrid approach.

4.3 Hyperparameter Setting

All models were fine-tuned using a consistent set of optimization hyperparameters, with minor adjustments made to accommodate model size and memory requirements. We trained DeBERTa-base, DeBERTa-large, RooseBERT-base-cased-cont, and BigBird-RoBERTa-base for 5 epochs, while an extended schedule of 10 epochs was additionally applied to DeBERTa-base and RooseBERT-base-cased-cont to improve convergence. The batch size per optimization step was set to 2 for DeBERTa-base, DeBERTa-large, and RooseBERT-base-cased-cont, whereas BigBird-RoBERTa-base was restricted to a batch size of 1 due to its higher memory footprint. To maintain effective batch sizes, we used gradient accumulation, with 8 accumulation steps for DeBERTa-base, DeBERTa-large, and RooseBERT-base-cased-cont, and 16 steps for BigBird-RoBERTa-base. Across all configurations, we used a learning rate of 10^{-5}, a warmup ratio of 0.1, and a maximum gradient norm of 1. Model selection was based on the macro F1-score computed on the development set, and the checkpoint achieving the highest score was retained as the final model.

5 Results and Discussion

Following our proposed methodology, we obtained an official evaluation macro F1-score of 0.61 on Subtask 2 (evasion-level classification) and 0.82 on Subtask 1 (clarity-level classification). As shown in Table 1, incorporating the top-5 evasion-level techniques predicted by the best encoder-only model as candidate labels in the LLM prompting strategy leads to improved performance compared to using either the best encoder-only model or the best LLM in a standalone classification setting.

Under five training epochs, both truncated DeBERTa-large and dropped DeBERTa-large achieve the best performance among all encoder-only models (Table 2). This indicates that larger models are particularly capable of capturing contextual information necessary for evasion technique classification. BigBird-RoBERTa-base, in contrast, underperforms substantially on the development set. This suggests that models designed for long-context input require more training epochs to effectively learn meaningful representations from the same number of observations. For RooseBERT-base, the difference in performance between 5 and 10 epochs is minimal. This implies that pretraining on a domain-specific corpus—here, political debates—enables the model to reach a local optimum, whereas DeBERTa-base continues to benefit from additional epochs.

Large language models, such as DeepSeek and Kimi-K2, outperform most encoder-only models in the evasion technique classification task (Table 3), and DeepSeek and Kimi-K2 perform comparably. This suggests that LLMs can achieve performance comparable to truncated DeBERTa-large even without labeled data.

Finally, combining the top-3 and top-5 predictions from the best encoder-only model as candidate labels for zero-shot and few-shot LLM inference yields the strongest results. Few-shot Kimi-K2 using the top-5 label set achieves a macro F1-score of 0.52 on Subtask 2 of SemEval Task 6, ranking 4th on the second shared subtask development

Preprocessing	Model	Acc@3	Acc@5	Acc	Prec	Recall	F1 (Sub 2)	F1 (Sub 1)
Training Epochs: 5								
Drop	DeBERTa-base	0.64	0.92	0.27	0.27	0.20	0.20	0.28
	DeBERTa-large	0.72	0.91	0.32	0.28	0.34	0.33	0.56
Truncate	DeBERTa-base	0.67	0.92	0.34	0.23	0.28	0.28	0.50
	DeBERTa-large	**0.77**	**0.97**	**0.39**	**0.43**	**0.43**	**0.46**	**0.65**
	RooseBERT-base-cased	0.68	0.93	0.28	0.36	0.31	0.31	0.43
Allow	BigBird-RoBERTa-base	0.67	0.94	0.28	0.42	0.20	0.21	0.43
Training Epochs: 10								
Truncate	DeBERTa-base	0.73	0.95	0.36	0.32	0.34	0.40	0.58
	RooseBERT-base-cased	0.70	0.93	0.31	0.43	0.35	0.33	0.44

Table 2: Model inference performance on Subtask 1 and Subtask 2 of the development set for encoder-only models under different preprocessing strategies. All metrics except macro F1-scores are for Subtask 2 which use annotator 3 as the reference label, while macro F1-scores are taken from the Codabench development phase shared task leaderboard.

Labels	Model	Method	F1 (S2)	F1 (S1)
All 9	DeBERTa-large	Fine-tuned	0.46	0.65
All 9	DeepSeek	Zero-shot	0.39	-
All 9	DeepSeek	Few-shot	0.44	-
All 9	Kimi-K2	Zero-shot	0.44	-
All 9	Kimi-K2	Few-shot	0.45	-
Top 3	Kimi-K2	Zero-shot	0.46	0.67
Top 3	Kimi-K2	Few-shot	0.50	**0.71**
Top 5	Kimi-K2	Zero-shot	0.47	0.67
Top 5	Kimi-K2	Few-shot	**0.52**	0.70

Table 3: Model inference performance on Subtask 1 and Subtask 2 of the development set. All macro F1-scores are from the Codabench development leaderboard.

Evasion Technique	Prec	Recall	F1	#
Claims ignorance	0.42	0.71	0.53	7
Clarification	1.00	1.00	1.00	4
Declining to answer	0.67	0.57	0.62	14
Deflection	0.17	0.30	0.22	23
Dodging	0.59	0.47	0.52	43
Explicit	0.53	0.78	0.63	80
General	0.48	0.25	0.33	65
Implicit	0.29	0.24	0.26	67
Partial/half-answer	1.00	0.20	0.33	5

Table 4: Our best hybrid system performance for each evasion technique evaluated on the development set.

phase leaderboard. This highlights the effectiveness of a hybrid approach where encoder-only models identify likely evasion techniques, and LLMs refine predictions by leveraging full input context.

Among the more frequent classes, performance varies considerably (Table 4). The model achieves its highest F1 on Explicit (0.63) and performs moderately on Dodging (0.52), while scores are much lower for General (0.33), Implicit (0.26), and especially Deflection (0.22). This suggests that some

System	F1 (S2)	F1 (S1)
Proposed Hybrid Encoder-LLM	**0.61**	**0.82**
Baseline Fine-tuned LLM	0.57	0.82

Table 5: System performance comparison on Codabench evaluation set between our proposed hybrid system and the baseline system proposed by Thomas et al. (2026) on Subtask 1 (S1) and Subtask 2 (S2) evaluation dataset using macro F1-score.

evasion types are inherently harder to separate. Explicit answers are easier because they tend to contain clear lexical signals that directly match the question, making the decision boundary more straightforward. Dodging is also relatively easier since it often involves a visible topic shift or a response that clearly ignores the question. In contrast, general and implicit replies rely on subtle pragmatic cues. They remain related to the question but avoid commitment through vagueness or indirectness, which makes them harder to distinguish both from each other and from partial or explicit answers.

As illustrated in Table 5, while both systems achieve parity in clarity classification (Subtask 1), our hybrid model outperforms the baseline in more granular evasion-level detection (Subtask 2). This performance gap suggests that the encoder-only constraint prediction space effectively guides the LLM toward more accurate evasion technique identification.

6 Conclusion

In this work for the SemEval-2026 Task 6 CLARITY Unmasking Political Interview, we undertook

a systematic investigation of deep learning models, focusing particularly on handling the significant long-context challenge present in the presidential interview QA dataset. We propose a hybrid inference approach that constrains the LLM prediction space using top-k candidate labels generated by our best-performing encoder-only model. This strategy proved superior, with the few-shot Kimi-K2 model using the top-5 label constraint achieving the top performance with a macro F1-score of 0.82 and 0.61 on the evaluation set, ranking 7th and 2nd on the first and second final shared subtask leaderboard.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Deborah Dore, Elena Cabrio, and Serena Villata. 2025. Roosebert: A new deal for political language modelling. *Preprint*, arXiv:2508.03250.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

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.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

A Prompting Details

Zero-shot prompt for classification: The following prompt was used for addressing the evasion-level classification in the zero-shot scenario.

```
system_prompt = "You are a text classification model. Based on a segment of the interview in which the interviewer poses a series of questions, choose ONLY the most appropriate type of response in Possible labels for the response provided.

Return results in the format:
SAMPLE_i_LABEL: <label>"

content_prompt = "You will classify multiple texts:
=== SAMPLE {i} ===

Text:
{text}

Possible labels:
{labels}

Output format:
SAMPLE_{i}_LABEL: <label>"
```

Few-shot prompt for classification: The following prompt was used for addressing the evasion-level classification in the few-shot scenario.

```
system_prompt = "You are a text classification model. Based on a segment of the interview in which the interviewer poses a series of questions, choose ONLY the most appropriate type of response in Possible labels for the response provided.

Return results in the format:
SAMPLE_i_LABEL: <label>"

content_prompt = "Here is one small example for each term of the label:

Question: Do you have your own views about PR at Westminster don't you?
Answer: I do.
Label: Explicit
Explanation: The answer directly gives the info requested.

Question: Are you going to watch television?
Answer: What else is there to do?
Label: Implicit
Explanation: They suggest planning to watch TV, despite not explicitly stating it.

Question: Do you like my new dress?
Answer: We are late.
Label: Dodging
Explanation: Does not even acknowledge the question and goes straight to another topic.

Question: Did you eat the last piece of pie?
Answer: I have to admit that this was a great recipe, I always like it when there are chocolate chips in the dough.
Label: Deflection
Explanation: Acknowledges the question but goes on a tangent about the chips, without answering.

Question: Did you enjoy the film?
Answer: The directing was great.
Label: Partial/half-answer
Explanation: Directing is only part of what constitutes a film.

Question: What's your favorite film?
Answer: Fight Club, Filth, and Hereditary.
Label: General
Explanation: The reply gives three movies instead of one, which makes the desired information unclear.

Question: The hypothesis I was discussing, wouldn't you regard that as a defeat?
Answer: I am not going to prophesy what will happen.
Label: Declining to answer
Explanation: Directly stating they won't answer.

Question: On what precise date did the government order the refit of the HMAS Kanimbla in preparation for its forward deployment to a possible war against Iraq?
Answer: I do not know that date. I will find out and let the House know.
Label: Claims ignorance
Explanation: Claims/admits they don't have the information.

Question: Was it your decision to release the fund?
Answer: You mean the public fund?
Label: Clarification
Explanation: Gives no data, asks for clarification.

You will classify multiple texts:
=== SAMPLE {i} ===

Text:
{text}

Possible labels:
{labels}

Output format:
SAMPLE_{i}_LABEL: <label>"
```

B Training Time

As indicated in Table 6, the computational overhead required to train the BigBird-RoBERTa-base

Preprocessing	Model	Time To Train
Training Epoch: 5 Epochs		
Drop	DeBERTa-base	18:52
	DeBERTa-large	58:55
Truncate	DeBERTa-base	30:12
	DeBERTa-large	1:24:43
	RooseBERT-base-cased	15:37
Allow	BigBird-RoBERTa-base	3:48:29
Training Epoch: 10 Epochs		
Truncate	DeBERTa-base	1:00:27
	RooseBERT-base-cased	32:02

Table 6: Training time using Google Colab L4 GPU comparison across different preprocessing methods and model architectures for 5 and 10 epochs.

Evasion Technique	Correct		Incorrect		#
	$\in top5$	$\notin top5$	$\in top5$	$\notin top5$	
Claims ignorance	5	0	2	0	7
Clarification	4	0	0	0	4
Declining to answer	8	0	4	2	14
Deflection	7	0	16	0	23
Dodging	20	0	23	0	43
Explicit	61	1	17	1	80
General	16	0	49	0	65
Implicit	16	0	51	0	67
Partial/half-answer	0	1	0	4	5

Table 8: Total counts of each evasion technique being predicted by our hybrid approach aggregated by the correctness of the system's predictions and the availability of evasion technique as constraint labels in LLM prompt.

model over five epochs significantly exceeded that of all other experimental configurations. Despite this increased training duration, the model's performance remained inferior to the alternative architectures, surpassing only the dropped DeBERTa-base baseline, as shown in Table 2. Consequently, given the prohibitive training latency coupled with suboptimal predictive accuracy, the BigBird-RoBERTa-base model was excluded from further iterative development.

C Qualitative Error Analysis

Evasion Technique	Acc@5	#
Claims ignorance	1.00	7
Clarification	1.00	4
Declining to answer	0.86	14
Deflection	1.00	23
Dodging	1.00	43
Explicit	1.00	80
General	1.00	65
Implicit	0.98	67
Partial/half-answer	0.00	5

Table 7: Accuracy@5 model performance of our best encoder-only model (truncated DeBERTa-large) for each evasion technique evaluated on the development set.

As illustrated in Table 7, the high-performing encoder-only model successfully identified the relevant evasion techniques within the top-5 candidates for nearly all instances. However, a notable exception was observed regarding the partial/half-answer technique, which the model failed to detect entirely across the dataset. This illustrates upper bound performance of our hybrid BERT-LLM approach in evasion-level classification.

Conversely, the experimental results detailed in Table 4 and Table 7 demonstrate that the proposed hybrid architecture effectively identifies partial/half-answers. This capability persists even when the technique is omitted from the prediction prompt's constraint labels, which were generated by the standalone encoder model. Further investigation into these dynamics, summarized in Table 8, reveals two successful and seven failed instances where the Large Language Model (LLM) predicted evasion-label classes—specifically declining to answer, explicit and partial/half-answer—despite their absence from the initial prediction constraints.