Duluth at SemEval-2026 Task 6: DeBERTa with LLM-Augmented Data for Unmasking Political Question Evasions

Shujauddin Syed & Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812 USA
{syed0093, tpederse}@d.umn.edu

Abstract

This paper presents the Duluth approach to SemEval-2026 Task 6 on CLARITY: Unmasking Political Question Evasions. We address Task 1 (clarity-level classification) and Task 2 (evasion-level classification), both of which involve classifying question–answer pairs from U.S. presidential interviews using a two-level taxonomy of response clarity. Our system is based on DeBERTa-V3-base, extended with focal loss, layer-wise learning rate decay, and boolean discourse features. To address class imbalance in the training data, we augment minority classes using synthetic examples generated by Gemini 3 and Claude Sonnet 4.5. Our best configuration achieved a Macro F1 of 0.76 on the Task 1 evaluation set, placing 8th out of 40 teams. The top-ranked system (TeleAI) achieved 0.89, while the mean score across participants was 0.70. Error analysis reveals that the dominant source of misclassification is confusion between Ambivalent and Clear Reply responses, a pattern that mirrors disagreements among human annotators. Our findings demonstrate that LLM-based data augmentation can meaningfully improve minority-class recall on nuanced political discourse tasks.

1 Introduction

The SemEval 2026 Task 6 on CLARITY: Unmasking Political Question Evasions (Thomas et al., 2026), challenges participants to automatically detect and classify evasive communication strategies in political discourse.

The task is organized around a two-level taxonomy of response clarity (Thomas et al., 2024): Task 1 requires classifying question–answer pairs from U.S. presidential interviews into three clarity categories (Clear Reply, Ambivalent, Clear Non-Reply), while Task 2 targets nine fine-grained evasion techniques.

Our system[1] is built on DeBERTa-V3-base (He et al., 2023), extended with focal loss (Lin et al., 2017), layer-wise learning rate decay, and two boolean discourse features extracted from the original data. To address severe class imbalance in the training set (59% Ambivalent, 31% Clear Reply, 10% Clear Non-Reply), we augmented the minority classes using synthetic examples generated by Gemini 3[2] and Claude Sonnet 4.5[3].

Our Gemini-augmented system achieved a Macro F1 of 0.76 on the evaluation set for Task 1, placing 8th out of 40 teams. Error analysis reveals that the primary challenge lies in distinguishing Ambivalent responses from Clear Replies, a confusion that mirrors disagreements observed among human annotators.

2 Task Description

This task addresses the computational detection and classification of evasive communication strategies in political discourse. The task is grounded in well-established theories of equivocation from political science (Bavelas et al., 1988; Bull, 1994; Bull and Strawson, 2019), which show that politicians provide clear responses to only 39–46% of questions during televised interviews. The shared task comprises two subtasks applied to question–answer (QA) pairs extracted from U.S. presidential interviews, organized around a two-level hierarchical taxonomy of response clarity proposed by Thomas et al. (2024).

Task 1 – Clarity-level Classification: Given a question–answer pair from a political interview, classify the response into one of three clarity categories:

Clear Reply: *The requested information is explicitly provided.*

Ambivalent: *A response is given but allows for multiple interpretations (e.g., implicit, general,*

[1] Our code is publicly available at our GitHub repository.

[2] See the Gemini 3 model card.

[3] See the Claude Sonnet 4.5 model card.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 650–657
July 3-4, 2026

partial, or deflective answers).

Clear Non-Reply: *The respondent openly refuses to share information, claims ignorance, or requests clarification.*

Formally, let q denote a question and a its corresponding answer. The task is to learn a function $f_1 : (q, a) \rightarrow \{1, 2, 3\}$ mapping each QA pair to one of the three clarity labels. This is a single-label, multi-class classification task.

Task 2 – Evasion-level Classification: Given the same question–answer pair, classify the response into one of 9 fine-grained evasion techniques that form the lower level of the taxonomy. These techniques are grouped under the three clarity categories as follows: *Explicit* (Clear Reply); *Implicit*, *General*, *Partial*, *Dodging*, and *Deflection* (Ambivalent); and *Declining to answer*, *Claims ignorance*, and *Clarification* (Clear Non-Reply).

The task is to learn a function $f_2 : (q, a) \rightarrow \{1, \ldots, 9\}$, mapping each QA pair to one of the nine evasion labels. This is also a single-label, multi-class classification task.

Our submission addresses both tasks. Our primary leaderboard submission targets Task 1, while we additionally explore Task 2 to investigate whether fine-grained evasion classification can inform and improve clarity-level predictions.

3 Related Work

Recent NLP research has begun to operationalize political discourse analysis building on the theoretical frameworks of political equivocation. Ferracane et al. (2021) crowdsourced annotations of political interview answers to capture subjective judgments about whether respondents intended to answer and whether their responses were truthful. In contrast, Thomas et al. (2024) introduced the QEvasion dataset and a two-level taxonomy that focuses on the *clarity* of responses rather than speaker intent, showing that fine-grained evasion labels can improve high-level clarity classification. Their work forms the basis of the CLARITY shared task and the dataset used in this paper.

Modeling such distinctions requires architectures capable of capturing subtle linguistic patterns while handling the inherent class imbalance in political discourse data. Transformer-based encoders such as DeBERTa-V3 (He et al., 2023) offer strong representational power through their disentangled attention mechanism. To address the severe imbalance between clear replies and evasive responses, we draw on focal loss (Lin et al., 2017), which reweights the training signal toward hard examples, and data augmentation strategies such as EDA (Wei and Zou, 2019) and context-aware synthetic generation (Park et al., 2024), both of which have been shown to improve minority-class performance. We additionally employ layer-wise learning rate decay (Zhang et al., 2021) to preserve pretrained knowledge in lower layers while allowing task-specific adaptation in higher layers. Our system integrates these techniques to address both the linguistic complexity and data imbalance inherent in the CLARITY task.

4 System Overview

This section describes the system we submitted to the leaderboard, including data augmentation, model architecture, and training details.

4.1 Data Augmentation for Class Imbalance

The QEvasion training set (Thomas et al., 2024) is imbalanced: Ambivalent (59.2%, 2,040), Clear Reply (30.5%, 1,052), and Clear Non-Reply (10.3%, 356). Preliminary experiments showed that this imbalance led to poor minority-class recall (Clear Non-Reply F1 < 0.40). To mitigate it, we augmented the training data using two external large language models, Gemini 3 and Claude Sonnet 4.5, which generated synthetic examples without accessing the test set.

Context-Aware Synthetic Generation (CASA) (Park et al., 2024):

Gemini 3 extracted rhetorical frames from minority classes (e.g., "I cannot comment on ongoing diplomatic discussions" for Clear Non-Reply) and combined them with randomized political contexts, producing 2,672 synthetic examples and perfectly balancing all classes at 2,040 each (6,120 total).

Lexical Paraphrasing (EDA-inspired) (Wei and Zou, 2019):

Claude Sonnet 4.5 applied four operations to answer texts (synonym replacement, random insertion, random swap, random deletion, $p = 0.1$), generating 1,086 synthetic examples and yielding a partially balanced distribution (Clear Reply: 1,498; Clear Non-Reply: 996; total 4,534).

We manually inspected 50 random samples to verify quality and political register. During training, we used a confidence-weighted loss (0.5× for Claude, 0.7× for Gemini) to reduce overfitting.

4.2 Model Selection: Why DeBERTa?

Before settling on DeBERTa, we experimented with several transformer architectures, including DistilBERT, BERT, and Political DEBATE. While BERT provided a strong baseline (0.56 test F1), we observed that DeBERTa-V3 and its variants consistently outperformed other models of comparable size on the development set. Its enhanced attention mechanism and improved pre-training led to a better understanding of nuanced political language. Preliminary runs with DeBERTa-V3-base achieved 0.64 dev F1 without any augmentation, leading us to adopt it as the foundation for our final system.

4.3 Final Model Architecture

Our leaderboard system is based on `microsoft/deberta-v3-base`[4] (184M parameters). We extended it with several enhancements:

Boolean features: Two statistically significant binary features from the original data, `affirmative_questions` and `multiple_questions`, are passed through a small feature processor (Linear→ReLU→Dropout) and concatenated with the pooled transformer output before classification.

Layer-wise Learning Rate Decay (LLRD): Lower layers receive progressively smaller learning rates, scaled by α^k with $\alpha = 0.9$, to preserve general knowledge while adapting higher layers.

Focal Loss (Lin et al., 2017): To focus on hard examples, we use

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \qquad (1)$$

where p_t is the model's estimated probability for the true class, α_t is the inverse-frequency class weight, and $\gamma = 2.0$.

Gradient accumulation: 4 steps simulate an effective batch size of 32.

Cosine annealing scheduler: 15% warmup steps followed by cosine decay.

Training runs for up to 6 epochs with early stopping (patience 3) based on validation Macro F1. Detailed hyperparameter tuning is described in Appendix A.

5 Experiments & Results

5.1 Evaluation Setup

The QEvasion dataset provides a predefined training set (3,448 samples) and test set (308 samples). We further split the training set into training (80%, 2,758) and development (20%, 690) using stratified sampling. Macro F1 is the official metric.

5.2 Leaderboard System Performance

We trained three variants of DeBERTa-V3-base on different data configurations (Section 4.1). Their performance on the test and evaluation datasets is shown in Table 2. The Gemini-augmented model achieved the highest test score and was our primary submission.

For Subtask 2, our DeBERTa-V3-large model with focal loss achieved a Macro F1 of 0.45 on the development phase (rank 9 of 24 teams), but dropped sharply to 0.28 on the evaluation phase (rank 30 of 33 teams). We attribute this drop to distribution shift between phases; a detailed analysis is provided in Appendix D.

For Subtask 1, our Gemini-augmented system achieved a test Macro F1 of 0.76, placing us **8th out of 40 teams** on the leaderboard. The top system scored 0.89, while the mean score was 0.70. This puts our system in the top 20% of participants, demonstrating the effectiveness of our augmentation and modeling choices.

5.3 Comparison with Baselines

Table 6 summarizes the performance of our best system against several baselines (detailed in Appendix B). The Gemini-augmented DeBERTa-V3-base outperforms all classical and simple transformer models. We also include a trivial baseline that always predicts the majority class (Ambivalent); its Macro F1 of 0.27 reflects the difficulty of the task and confirms that our models are learning meaningful patterns.

5.4 Error Analysis

To better understand the remaining errors, we examined confusion matrices for both the test set (308 samples) and the final evaluation set (237 samples) in Tables 3 and 4. The matrices reveal consistent patterns across both splits, suggesting our model generalizes well without overfitting to specific data characteristics.

The test set confusion matrix shows that the main source of confusion is between Ambivalent and Clear Reply, accounting for 68% of all errors (58 + 23 = 81 out of 119 total errors). This pattern persists in the evaluation set, where Ambivalent–Clear Reply confusion represents 65% of errors (20 + 16 = 36 out of 55 total errors). This consistent pattern

[4] See the DeBERTa-V3-base model page.

Table 1: Processing pipeline for an example question–answer pair. The example shows how raw input is transformed before being fed to the model.

Input	
Question	Will you increase funding for education?
Answer	I cannot comment on budget discussions at this time.
Ground truth	Clear Non-Reply
After formatting	Question: Will you increase funding for education? [SEP] Answer: I cannot comment on budget discussions at this time.
Boolean features	`affirmative_questions`=1, `multiple_questions`=0
Model output	Clear Non-Reply

Table 2: Performance of DeBERTa-V3-base variants

Configuration	Test F1	Eval F1
Original only	0.64	0.69
Claude-augmented	0.65	0.74
Gemini-augmented (submitted)	**0.66**	**0.76**
Top system (TeleAI)	–	**0.89**

Table 3: Confusion matrix for Gemini-augmented DeBERTa-V3-base on the test set (308 samples). Rows = true labels, columns = predictions. Row and column totals include percentages.

	Amb	Clear	Clear-N	**R-Total**
Amb	136	58	12	**206** (66.9%)
Clear	23	53	3	**79** (25.6%)
Clear-N	6	0	17	**23** (7.5%)
C-Total	**165** (53.5%)	**111** (36.1%)	**32** (10.4%)	**308**

indicates that the model fundamentally struggles with distinguishing partially informative answers from fully direct ones, even after data augmentation.

The model achieves a Macro F1 of 0.6364 on the test set, with per-class F1 scores of 0.73 (Ambivalent), 0.56 (Clear Reply), and 0.62 (Clear Non-Reply). Notably, the model perfectly distinguishes Clear Non-Reply from Clear Reply in both splits (zero false positives in the Clear Reply column for true Clear Non-Reply samples), though it still misclassifies some Clear Non-Reply instances as Ambivalent.

Appendix C presents multiple examples of each error type, revealing recurring patterns:

Ambivalent → Clear Reply errors: Answers contain both a direct statement and hedging (e.g., "We have increased funding, but we need to study the impact further"). The model latches onto the concrete claim while missing the qualifiers.

Table 4: Confusion matrix for Gemini-augmented DeBERTa-V3-base on the evaluation set (237 samples). Rows = true labels, columns = predictions.

	Amb	Clear	Clear-N	**R-Total**
Amb	91	20	6	**117** (49.4%)
Clear	16	68	1	**85** (35.8%)
Clear-N	12	0	23	**35** (14.8%)
C-Total	**119** (50.3%)	**88** (37.1%)	**30** (12.6%)	**237**

Clear Reply → Ambivalent errors: Answers are direct but use tentative or conditional language (e.g., "I believe we will consider raising funds"). The hedging vocabulary misleads the model into classifying them as Ambivalent.

Clear Non-Reply errors: Answers are evasive but contain factual statements or appear superficially responsive (e.g., "The budget is complex; I cannot comment now, but here are last year's numbers"). The inclusion of concrete facts confuses the classifier.

These observations suggest that further gains could be obtained by better modeling hedging and partial answers, perhaps through multi-task learning that jointly predicts both clarity and the presence of hedging language, or by incorporating external knowledge about political discourse patterns.

6 Additional Models

We also experimented with two domain-adapted or larger models; their results are summarized in Table 5.

6.1 Political DEBATE

`mlburnham/Political_DEBATE_base_v1.0`[5] is a DeBERTa-base model further pre-trained on political discourse. We fine-tuned it with the same advanced pipeline as DeBERTa-V3-base (LLRD, gra-

[5]See the Political DEBATE model page.

Table 5: Additional model results (development set)

Model	Test F1
Political DEBATE	0.57
DeBERTa-V3-large-NLI	0.66

dient accumulation, cosine annealing, early stopping), using learning rate 3e-5 and effective batch size 32.

6.2 DeBERTa-V3-large-NLI

`MoritzLaurer/DeBERTa-v3-large-nli`[6] (435M parameters) is pre-trained on multiple NLI datasets. Training followed the DeBERTa-V3-base pipeline but with batch size 4 and gradient accumulation steps increased to 8 (effective batch size 32). Although this model achieved comparable performance to our Gemini-augmented DeBERTa-V3-base, we selected the latter as our primary submission due to its lower training cost and faster iteration time.

6.3 Additional Results

Table 5 shows the performance of the additional models on the development set (test scores were not submitted).

7 Future Work and Conclusions

We presented an augmented DeBERTa-V3-base system for classifying the clarity of political interview responses. Our primary contribution is demonstrating that LLM-based synthetic data augmentation, combined with focal loss and careful hyperparameter tuning, can substantially improve minority-class performance on this task, raising evaluation Macro F1 from 0.69 (no augmentation) to 0.76 (Gemini-augmented). Participating in the CLARITY task reinforced that the core difficulty lies not in identifying outright non-replies, which our model handles well, but in the gray zone between Ambivalent and Clear Reply responses, where even human annotators disagree ($\kappa = 0.65$).

If we were to continue this work, we would pursue three directions.

First, a multi-task learning setup that jointly predicts clarity labels and evasion-level labels could leverage the hierarchical structure of the taxonomy, as prior work has shown that fine-grained evasion classification improves clarity-level predictions (Thomas et al., 2024).

[6]See the DeBERTa-v3-large-NLI model page.

Second, hedging-aware features such as explicit detection of qualifiers, conditionals, and topic shifts could help the model distinguish partially informative answers from genuinely direct ones, addressing the dominant error pattern in our system.

Third, our Subtask 2 results (Appendix D) exposed a critical need for more robust validation and minority-class augmentation at the nine-class level, where distribution shift between development and evaluation phases caused a steep performance drop.

8 Acknowledgments

The authors would like to thank the organizers for the opportunity to participate in SemEval-2026 Task 6 and the Department of Computer Science, University of Minnesota Duluth, for helping us with all resources needed to participate in this task. We also appreciate the valuable input and guidance provided by the reviewers.

References

Janet Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. 1988. Political equivocation: A situational explanation. *Journal of Language and Social Psychology*, 7:137–145.

Peter Bull. 1994. On identifying questions, replies, and non-replies in political interviews. *Journal of Language and Social Psychology*, 13:115–131.

Peter Bull and Will Strawson. 2019. Can't answer? won't answer? an analysis of equivocal responses by theresa may in prime minister's questions. *Parliamentary Affairs*, 73.

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, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *CoRR*, abs/1708.02002.

Chanwoo Park and 1 others. 2024. Casa: Context-aware synthetic augmentation for text classification. *arXiv preprint arXiv:2403.02990*.

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.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning. *Preprint*, arXiv:2006.05987.

A Hyperparameter Tuning Details

We performed a focused hyperparameter search on the original training set, fixing gradient accumulation at 4 and $\gamma = 2.0$. We varied the learning rate and LLRD factor:

- Learning rates tested: 2e-5, 3e-5, 5e-5 (with $\alpha = 0.9$). The best was 3e-5.
- With learning rate 3e-5, we explored $\alpha = 0.8, 0.9, 0.95$. $\alpha = 0.9$ gave the most stable training.

These values were used for all subsequent experiments, including the augmented-data variants.

B Baseline Models

Here we describe the configurations of models that served as baselines.

B.1 Classical Machine Learning Models

TF-IDF + Logistic Regression – scikit-learn's `TfidfVectorizer` with `max_features=5000`, `ngram_range=(1,2)`, `min_df=2`, `max_df=0.95`, `stop_words='english'`. Classifier: LogisticRegression(`class_weight='balanced'`, `C=1.0`).

SVM – Same TF-IDF with L2 normalization and sublinear tf scaling. Grid search over `C=[0.1,1.0,10.0]`, kernels (`linear`,`rbf`), 3-fold CV, `class_weight='balanced'`.

Random Forest – 100 trees, `max_depth=20`, `min_samples_split=10`, `min_samples_leaf=4`, `class_weight='balanced'`.

Table 6: Comparison with baseline models

Model	Test F1
Majority class (Ambivalent)	0.2700
TF-IDF + Logistic Regression	0.4476
SVM (linear)	0.4270
Random Forest	0.4256
DistilBERT	0.5158
BERT-base	0.5628
DeBERTa-V3-base (Gemini)	**0.66**

B.2 Simple Transformers

All transformers were fine-tuned with HuggingFace Transformers using AdamW (weight decay 0.01), gradient clipping (1.0), linear warmup (10%) and linear decay, and class-weighted cross-entropy.

DistilBERT – `distilbert-base-uncased`[7], 4 epochs, batch size 16, learning rate 2e-5.

BERT – `bert-base-uncased`[8], 4 epochs, batch size 8 (due to memory), learning rate 2e-5.

C Detailed Error Examples

This appendix provides three representative examples for each of the main error types observed in the test set predictions of our Gemini-augmented DeBERTa-V3-base model. The examples illustrate the recurring patterns that cause misclassifications.

C.1 Ambivalent → Clear Reply Errors (20 total)

The model over-commits to a label when the response sounds direct but does not fully address the question.

- **Example 1 (June 2017):**
  Question: "So he said those things under oath?"
  Answer: "One-hundred percent. I didn't say under oath—I hardly know the man... No, I didn't say that, and I didn't say the other."
  Analysis: The opening "One-hundred percent" signals directness, but the answer neither confirms nor denies the claim—it pivots to what the speaker did not say. The model latches onto the confident opening and ignores the subsequent hedging.

[7] See the DistilBERT model page.
[8] See the BERT-base model page.

- **Example 2 (February 2016):**
  Question: "Would you consider a recess appointment if your nominee is not granted a hearing?"
  Answer: "I think that we have more than enough time to go through regular order... I expect them to hold hearings. I expect there to be a vote."
  Analysis: The response never actually answers the yes/no question about a recess appointment. The model is fooled by the confident declarative tone and the detailed discussion of the regular process.

- **Example 3 (June 2006):**
  Question: "Do you have a specific target for how much you want the violence reduced?"
  Answer: "Enough for the Government to succeed."
  Analysis: The answer sounds direct but is entirely non-specific. The question asks for a numerical target, and the response gives a vague criterion. The model mistakes this for a Clear Reply because of its assertive phrasing.

C.2 Clear Reply → Ambivalent Errors (16 total)

The model under-commits when responses are brief, blunt, or terse—mistaking conciseness for hedging.

- **Example 1 (January 1953):**
  Question: "You have described it as a billion dollar steal?"
  Answer: "You left off two zeros. It's a hundred billion dollars."
  Analysis: This is a crisp, direct correction. The model likely penalizes the brevity or the indirectness of the corrective framing (a question–answer that corrects without a simple "yes").

- **Example 2 (September 1980):**
  Question: "Does an apology rule out the question of honor?"
  Answer: "Yes. The United States is not going to apologize."
  Analysis: An unambiguous "Yes" followed by a clear statement. The longer elaboration after the "Yes" may have confused the model into interpreting the response as more complex than a simple affirmation.

- **Example 3 (July 2022):**
  Question: "So you don't expect to bring up human rights?"
  Answer: "I will bring up—I always bring up human rights. I always bring up human rights."
  Analysis: A direct contradiction of the premise. The repetition and the interjection from another speaker (visible in the transcript) may have introduced spurious signals of ambiguity, causing the model to miss the clear reply.

C.3 Clear Non-Reply → Ambivalent Errors (12 total)

The model fails to recognize explicit refusals or deflections as non-replies, possibly because the responses contain some substantive-sounding content or are too short.

- **Example 1 (July 2017):**
  Question: "And about his role in Syria and the region?"
  Answer: "Whose role?"
  Analysis: A pure deflection—answering a question with a question. Very short responses may not provide enough signal for the model to confidently classify them as non-replies.

- **Example 2 (August 1992):**
  Question: "But they were ready to move sooner if asked, weren't they?"
  Answer: "I'm not going to go into that because... what you seem to be interested in is kind of assigning blame. That is not what's at stake here, and I don't want to participate in that."
  Analysis: An explicit refusal to answer—but the model may have read the surrounding substantive content (the discussion of blame) as partial engagement, leading to an Ambivalent prediction.

- **Example 3 (January 1953):**
  Question: "Mr. President, there is one question that is left unanswered."
  Answer: "What's that?"
  Analysis: Another question-as-response, deflecting entirely. The extreme brevity leaves the model without enough context to detect the non-reply pattern.

D Subtask 2 Evaluation Phase Analysis

Our Subtask 2 system exhibited a substantial performance gap between the development phase (Macro F1 = 0.45, rank 9/24) and the evaluation phase (Macro F1 = 0.28, rank 30/33). Post-hoc analysis identified the following:

We identified a flaw in validation during the development phase: the original training script assigned dummy labels (all zeros) to the held-out test split, meaning early stopping and model selection were optimized against meaningless validation metrics. The model checkpoint saved as "best" was effectively selected at a random epoch. Despite retraining with a proper stratified cross-validation split, which raised internal validation F1 to 0.51, the evaluation phase score remained at 0.28, suggesting substantial distribution shift between the development and evaluation sets. The nine-class imbalance (13.3× ratio between the largest and smallest class) likely exacerbated this shift, as minority classes (*Partial/half-answer*: 79 training samples, *Clarification*: 92) are highly sensitive to domain variation. Future work should address this through cross-domain validation and more aggressive minority-class augmentation.