SG-UniBuc-NLP at SemEval-2026 Task 6: Multi-Head RoBERTa with Chunking for Long-Context Evasion Detection

Gabriel Stefan and Sergiu Nisioi
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
gabrielstefan04@gmail.com, sergiu.nisioi@unibuc.ro

Abstract

We describe our system for SemEval-2026 Task 6 (CLARITY: Unmasking Political Question Evasions), which classifies English political interview responses by coarse-grained clarity (3-way) and fine-grained evasion strategy (9-way). Since responses frequently exceed the 512-token limit of standard Transformer encoders, we apply an overlapping sliding-window chunking strategy with element-wise Max-Pooling aggregation over chunk representations. A shared RoBERTa-large encoder supplies two task-specific heads trained jointly via a multi-task objective, with inference-time ensembling over 7-fold stratified cross-validation. Our system achieves a Macro-F1 of 0.80 on Subtask 1 and 0.51 on Subtask 2, ranking 11th in both subtasks.

1 Introduction

Political interviews are a primary venue of democratic accountability, yet politicians routinely avoid giving direct answers. A meta-analysis of five televised interview studies by Bull (2003) found that politicians provided clear responses to only 39–46% of questions, compared to 70–89% in non-political settings. This phenomenon, widely termed equivocation or evasion in the social-science literature (Harris, 1991; Bull, 2003; Rasiah, 2010), encompasses a range of rhetorical strategies, from subject shifts and partial answers to deliberate ambiguity, and has been the subject of detailed typological study (Clayman, 2001; Bull and Mayer, 1993). Despite this social-science foundation, the automatic detection of such strategies has received limited attention in NLP (Thomas et al., 2024). SemEval-2026 Task 6 (*CLARITY*: *Unmasking Political Question Evasions*; Thomas et al. 2026) addresses this gap by framing clarity and evasion detection as a supervised classification task over English political interview question–answer pairs, requiring systems to predict both a coarse-grained three-way *Clarity* label and a fine-grained nine-way *Evasion* label.

A central challenge is that political responses frequently exceed the 512-token input limit of standard Transformer encoders (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019): naïve truncation risks discarding the precise span where evasion cues appear. We therefore segment each question–answer pair into overlapping 512-token chunks with a stride of 256 tokens, encode each chunk independently with a shared RoBERTa encoder, and aggregate the resulting chunk representations via element-wise Max-Pooling to obtain a single response-level vector (Pappagari et al., 2019). Two task-specific linear heads, a 3-way Clarity classifier and a 9-way Evasion classifier, are jointly trained on top of this shared encoder via a combined cross-entropy objective (Caruana, 1997; Ruder, 2017), allowing the coarser clarity signal to regularize the more challenging evasion classification task. We train the entire system using 7-fold stratified cross-validation and combine the resulting fold models by averaging predicted probabilities at inference time (Galar et al., 2012).

On the official SemEval-2026 Task 6 evaluation, our ensemble achieves a Macro-F1 of 0.80 on Subtask 1 (Clarity), ranking 11th of 41 teams, and a Macro-F1 of 0.51 on Subtask 2 (Evasion), ranking 11th of 33 teams. Error analysis reveals two recurring failure modes driven by class imbalance and semantic overlap: *Ambivalent* acts as a classification sink in Subtask 1, while performance on Subtask 2 severely degrades for minority categories with fine-grained pragmatic distinctions. Our implementation is publicly available.[1]

[1] https://github.com/gabriel-stefan/political-evasion-detection

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 964–972
July 3-4, 2026

2 Background

2.1 Task and Data

SemEval-2026 Task 6 (*CLARITY*; Thomas et al. 2026) frames political evasion detection as a supervised classification problem over English question–answer (QA) pairs. Each instance consists of a question Q and a response A; systems predict two labels simultaneously. **Subtask 1** requires 3-way *Clarity* classification: *Clear Reply* (question fully addressed), *Clear Non-Reply* (speaker explicitly refuses), or *Ambivalent* (response admits multiple interpretations). **Subtask 2** requires 9-way *Evasion* classification into the taxonomy leaves of Thomas et al. (2024): *Explicit*, *Dodging*, *Implicit*, *General*, *Deflection*, *Partial/half-answer*, *Clarification*, *Claims ignorance*, or *Declining to answer*. We participated in both subtasks; performance is evaluated with Macro-F1 (Thomas et al., 2026).

Example:
Q: *Based on your long experience, how does that change Finland's place in the world?*
A: *Well, first of all, the context in which I said that was: The gentleman who occupies a seat...*
Labels: *Ambivalent* (Subtask 1), *Dodging* (Subtask 2).

The CLARITY dataset (Thomas et al., 2024) comprises 3,756 English QA pairs from 287 official White House interview transcripts (2006–2023), split into 3,448 training and 308 development instances (Thomas et al., 2026). Label distributions are substantially skewed: *Ambivalent* constitutes 59% of training instances, while *Clear Non-Reply* accounts for only 10%. At the evasion level, *Explicit* (31%) and *Dodging* (21%) dominate, while *Partial/half-answer* (2%), *Clarification* (3%), and *Claims ignorance* (3%), among others, are severely underrepresented. Inter-annotator agreement stands at Fleiss $\kappa = 0.64$ for Subtask 1 and $\kappa = 0.48$ for Subtask 2 (Thomas et al., 2024), underscoring that fine-grained evasion categorization is challenging even for human judges. For development and test instances, multiple annotator labels are provided, and a prediction is scored as correct if it matches any of them (Thomas et al., 2026).

2.2 Related Work

The CLARITY taxonomy builds on classical social-science typologies of political equivocation (Harris, 1991; Bull, 2003; Rasiah, 2010), consolidated into a computational label set by Thomas et al. (2024).

From a modeling perspective, the task poses a long-input classification challenge, as responses frequently exceed standard Transformer encoder limits. Dedicated long-context encoders such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2021) address this via sparse attention, and more recent encoder-only models extend context further still (Warner et al., 2025). An alternative line of work represents documents hierarchically by encoding fixed-length segments and aggregating their representations (Dietterich et al., 1997; Ilse et al., 2018; Pappagari et al., 2019), which preserves full-document evidence while remaining compatible with standard pretrained encoders.

3 System Overview

3.1 Hierarchical Input Processing

A key challenge in CLARITY is that responses can span thousands of tokens, exceeding the 512-token input limit of standard pretrained encoders such as RoBERTa (Liu et al., 2019). Since evasion cues may appear anywhere in the response, naïve truncation can remove critical evidence. We therefore adopt a sliding-window chunking strategy.

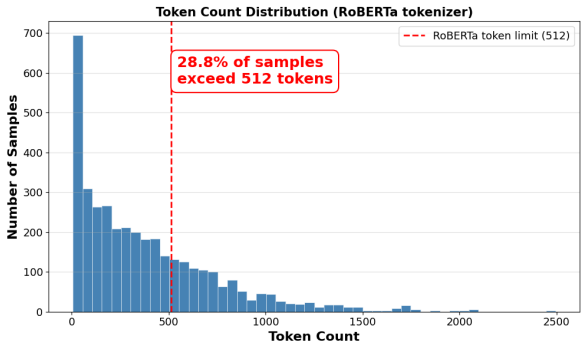

Figure 1: Distribution of token counts for the concatenated question–answer input sequence in the CLARITY dataset. The vertical red dashed line marks the 512-token limit of standard RoBERTa models.

We adopt hierarchical chunking over long-context architectures (e.g., Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2021)) primarily for memory efficiency, as chunking strictly bounds the memory footprint and prevents out-of-memory errors during fine-tuning. Additionally, this approach is straightforward to implement, as it integrates directly with any pretrained encoder without architectural modifications. Finally, our data distribution justifies this choice: since only 28.8% of instances exceed 512 tokens (Figure 1), the computational overhead of sparse-attention

mechanisms is unnecessary for most inputs.

Given a question Q and answer A, we construct a single input string using explicit prefixes: `Question: {Q}\nAnswer: {A}`. Let T denote the tokenized sequence of this full concatenation. Since $|T|$ can exceed the encoder limit, we segment T into overlapping windows of length $L = 512$ with stride $S = 256$:

$$C_k = T[kS : kS + L] \quad \text{for } k = 0, \ldots, M-1, \tag{1}$$

where $M = \left\lceil \frac{\max(|T|-L,0)}{S} \right\rceil + 1$ is the number of chunks for the instance. The final chunk always extends to $|T|$ and is padded to length L if shorter. The window length $L = 512$ is fixed by the encoder's maximum positional embedding capacity and is not a tunable hyperparameter. The stride $S = 256$ was set as a principled 50% overlap rather than through grid search: this overlap ensures that every token (except those in single-chunk documents) appears in at least two consecutive windows, reducing the risk that a chunk boundary splits a semantically coherent evasion span.

Each chunk C_k is encoded independently with a shared RoBERTa-large encoder. We extract the hidden state at position 0 of each chunk as a fixed-size chunk representation:

$$h_k = H_k[0, :] \in \mathbb{R}^d, \tag{2}$$

where $H_k \in \mathbb{R}^{L \times d}$ is the final hidden state matrix for chunk k and $d = 1024$ for RoBERTa-large. Note that only the first chunk (k=0) begins with the actual `<s>` special token; for subsequent chunks, position 0 corresponds to the first token of that window.

We then aggregate all chunk vectors for the same instance using element-wise Max-Pooling:

$$v_j = \max_{k=0}^{M-1} h_{k,j} \quad \text{for } j = 1, \ldots, d, \tag{3}$$

yielding a single response-level representation $v \in \mathbb{R}^d$.

Algorithm 1 Long-Input Encoding Pipeline: overlapping chunking, per-chunk RoBERTa encoding, and element-wise Max-Pooling aggregation into a single response vector v.

Input: token sequence T (length $|T|$), max length $L = 512$, stride $S = 256$
Output: response vector $v \in \mathbb{R}^d$
$chunks \leftarrow [\,]$
$start \leftarrow 0$
while $start < |T|$ **do**
  $end \leftarrow \min(start + L,\ |T|)$
  $C \leftarrow T[start : end]$
  pad C to length L and build attention mask
  $chunks.append(C)$
  if $end \geq |T|$ **then**
    break ▷ last chunk reaches end
  end if
  $start \leftarrow start + S$
end while
for all $C_k \in chunks$ **do** ▷ encoded in a single batched forward pass
  $H_k \leftarrow$ RoBERTa-large(C_k)
  $h_k \leftarrow H_k[0, :]$ ▷ position-0 embedding
end for
$v \leftarrow$ ElementWiseMax$(\{h_k\}_k)$
return v

3.2 Multi-Task Learning Heads

As shown in Figure 2, the pooled vector v (Eq. 3) is shared by both subtasks. We apply dropout (Srivastava et al., 2014) ($p = 0.1$) to v and feed it into two task-specific linear heads: a 3-way classifier for Clarity and a 9-way classifier for Evasion.

$$\hat{y}_c = \text{softmax}(W_c \cdot \text{Dropout}(v) + b_c), \tag{4}$$
$$\hat{y}_e = \text{softmax}(W_e \cdot \text{Dropout}(v) + b_e), \tag{5}$$

where $W_c \in \mathbb{R}^{3 \times d}$ and $W_e \in \mathbb{R}^{9 \times d}$.

We train both heads jointly using standard cross-entropy losses with equal weighting:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_e. \tag{6}$$

We investigate alternative loss functions in Section 5.

3.3 Training and Inference

We train using 7-fold stratified cross-validation, stratifying folds by the Subtask 1 (Clarity) labels to preserve class proportions. In each fold, we select the checkpoint that maximizes the combined validation score (defined below in Section 4.3), treating both subtasks equally. Full training hyperparameters are listed in Table 6.

At inference time, we ensemble all 7 fold models by averaging their predicted class probabilities and taking $\arg\max$:

$$\hat{y} = \arg\max_c \ \frac{1}{7} \sum_{i=1}^{7} p_c^{(i)}, \tag{7}$$

where $p_c^{(i)}$ is the predicted probability for class c from fold model i.

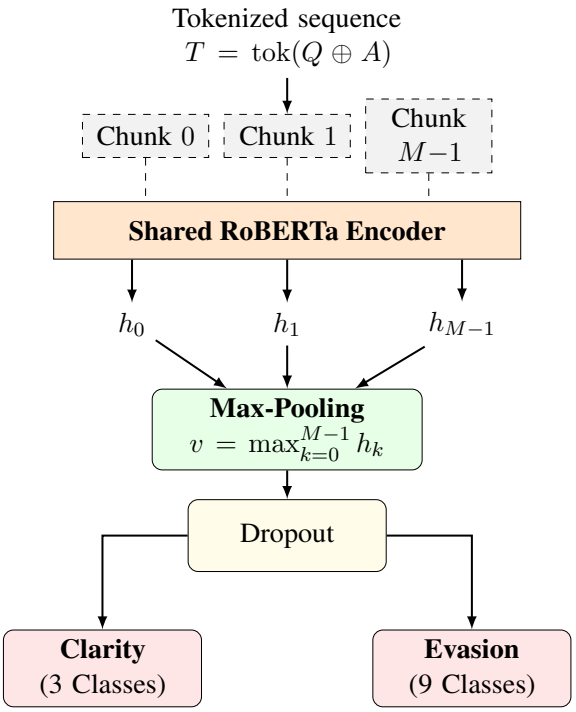

Figure 2: System architecture. The tokenized concatenated input ($Q \oplus A$) is split into overlapping chunks, each chunk is encoded by a shared RoBERTa encoder, chunk representations are aggregated via element-wise Max-Pooling, and two task-specific heads predict Clarity (3-way) and Evasion (9-way).

4 Experimental Setup

4.1 Data Splits

We follow the official SemEval-2026 Task 6 setup (Thomas et al., 2026). The organizers provide a labeled `train` split, a labeled `dev` split (released as `test` during the competition but used here as the official development set) and a blind `test` set. We train models on `train` and estimate performance via 7-fold stratified cross-validation (Section 3.3).

4.2 Preprocessing

Each instance is formatted as a single sequence using the template `Question: {Q}\nAnswer: {A}`. No truncation is applied prior to chunking; the full concatenated sequence is tokenized and passed directly to the sliding-window pipeline described in Section 3.1. Sequences are dynamically padded to the batch maximum length by the data collator; within the model, individual chunks shorter than L=512 tokens are zero-padded to L.

We fine-tune RoBERTa-large (Liu et al., 2019) using AdamW (Loshchilov and Hutter, 2019) with a linear learning-rate schedule and warmup; exact training hyperparameters and configuration details are provided in Appendix A.

4.3 Evaluation Metrics

Following the task guidelines (Thomas et al., 2026), the primary metric is Macro-F1 for both subtasks. Macro-F1 computes the unweighted average of per-class F1 scores. For early stopping and checkpoint selection during cross-validation, we compute the combined score:

$$\text{F1}_{\text{comb}} = \frac{1}{2}\left(\text{F1}_{\text{clarity}} + \text{F1}_{\text{evasion}}\right), \qquad (8)$$

which treats both subtasks equally and avoids optimizing for one at the expense of the other.

4.4 Implementation

We implement our system in Python 3.12 using PyTorch and Hugging Face Transformers (full dependencies provided in Appendix B). Random seeds are fixed to 42 for reproducibility. All experiments were conducted on a single NVIDIA RTX 3090 (24 GB VRAM). Training the full 7-fold ensemble takes approximately 5 hours.

5 Results and Analysis

5.1 Main Results

On the official SemEval-2026 Task 6 blind test set (submitted under the CodaBench username `gabriel-stefan`), our final ensemble achieves a Macro-F1 of 0.80 on Subtask 1 (Clarity), ranking 11th out of 41 teams. On Subtask 2 (Evasion), the system achieves a Macro-F1 of 0.51, ranking 11th out of 33 participating teams.

5.2 Ablation Studies

To isolate the impact of individual design choices, each ablation varies one component while keeping the rest fixed to our full configuration: hierarchical Max-Pooling, multi-task training, and 7-fold stratified cross-validation. We report mean and standard deviation across folds.

Pooling Strategy Table 1 shows Max-Pooling outperforms Mean-Pooling and the First-Chunk baseline. Mean-Pooling averages representations, potentially diluting localized signals. Conversely, element-wise Max-Pooling extracts the maximum feature activations across all chunks, creating a composite representation that likely preserves the strongest evasion cues regardless of position, explaining the improved fine-grained recall.

Pooling Method	Clarity F1	Evasion F1
First Chunk Only	0.67 ± 0.01	0.42 ± 0.01
Mean Pooling	0.68 ± 0.02	0.43 ± 0.02
Max-Pooling (Ours)	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.45 \pm 0.02}$

Table 1: Max-Pooling outperforms Mean-Pooling and the First-Chunk baseline across both subtasks (7-fold CV; mean $\pm$ std).

Multi-Task Learning Table 2 compares joint Multi-Task training against single-task models, demonstrating the benefit of auxiliary supervision. While Clarity performance remains constant, the Multi-Task objective improves Evasion Macro-F1 from 0.42 to 0.45. This suggests high-level clarity labels provide an effective regularization signal for the more complex evasion task.

Training Objective	Clarity F1	Evasion F1
Single-task (Clarity only)	$\mathbf{0.70 \pm 0.02}$	–
Single-task (Evasion only)	–	0.42 ± 0.01
Multi-task (Ours)	0.70 ± 0.02	$\mathbf{0.45 \pm 0.02}$

Table 2: Effect of multi-task learning (7-fold CV; mean $\pm$ std). Joint training improves the more difficult Evasion task via auxiliary regularization.

Ensemble Size To quantify the trade-off between robustness and computational cost, we trained three variants using stratified k-fold cross-validation with $k \in \{3, 5, 7\}$ (stratified by Subtask 1 labels). For each k, we report the mean $\pm$ standard deviation of validation Macro-F1 across the k held-out folds. At test time, each variant ensembles the k fold models by averaging their predicted probabilities. As shown in Table 3, increasing k yields consistent gains, especially for the more difficult Evasion subtask, at the cost of proportionally higher training and inference time.

Ensemble size	Clarity F1	Evasion F1	Rel. cost
3-fold	0.66 ± 0.01	0.42 ± 0.02	$3.0\times$
5-fold	0.68 ± 0.02	0.43 ± 0.03	$5.0\times$
7-fold (Ours)	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.45 \pm 0.02}$	$7.0\times$

Table 3: Increasing ensemble size consistently improves both subtasks at proportionally higher cost, with each larger k also providing more training data per fold model (7-fold CV; mean $\pm$ std).

Class Imbalance Mitigation To evaluate alternative strategies for handling class imbalance (Section 3.3), we trained two additional variants: inverse-frequency class-weighted cross-entropy and focal loss ($\gamma = 2$; Lin et al. 2017). Table 4 reports the 7-fold cross-validation results.

Loss function	Clarity F1	Evasion F1
Cross-entropy (ours)	0.70 ± 0.02	0.45 ± 0.02
Class-weighted CE	0.66 ± 0.02	0.41 ± 0.02
Focal loss ($\gamma = 2$)	0.69 ± 0.02	0.44 ± 0.02

Table 4: Effect of loss function on class imbalance (7-fold CV, mean $\pm$ std).

Neither alternative improves Macro-F1, but error patterns reveal a meaningful redistribution of performance rather than uniform degradation. Class weighting increases minority-class recall: *Clear Non-Reply* rises from 0.62 to 0.67, and *Partial/half-answer* increases from 0.00 to 0.10 on out-of-fold predictions (the only setting predicting it at all). However, these gains are eclipsed by majority-class precision drops (*Ambivalent* F1 falls from 0.78 to 0.75; *Explicit* F1 from 0.65 to 0.60), resulting in a net decrease in the equally-weighted Macro-F1. Focal loss falls between these extremes, avoiding sharp majority-class degradation but yielding only marginal minority-class recovery.

We attribute this to two compounding factors. First, data sparsity is the binding constraint: with only 79 *Partial/half-answer* instances, reweighting amplifies gradients but cannot compensate for lacking lexical and pragmatic diversity. The 10% recall confirms the model can detect this strategy, but the available signal remains too sparse and noisy. Second, dominant confusions reflect semantic overlap, not sampling bias. The largest error clusters involve pragmatically adjacent classes (*Implicit*, *Deflection*, *General*, *Dodging*) sharing surface features that confound even human annotators (Fleiss $\kappa = 0.48$). Reweighting amplifies gradients without resolving this representational ambiguity, potentially destabilizing the shared encoder with conflicting supervision for examples with similar surface forms. These findings suggest improving minority classes requires targeted data augmentation rather than loss engineering (Section 6).

5.3 Error Analysis

We analyze out-of-fold predictions on the 3,448 training instances (Table 5; confusion matrices in Appendix C).

Subtask 1 (Clarity) While Table 5 reports overall confidence, we observe that the model's mean prediction confidence remains remarkably high

Class	n	Acc	Conf
Subtask 1 (Clarity)			
Ambivalent	2040	0.784	0.936
Clear Non-Reply	356	0.643	0.927
Clear Reply	1052	0.623	0.935
Subtask 2 (Evasion)			
Clarification	92	0.707	0.892
Explicit	1052	0.643	0.830
Declining to answer	145	0.607	0.838
Claims ignorance	119	0.521	0.795
Dodging	706	0.474	0.757
General	386	0.306	0.710
Implicit	488	0.281	0.706
Deflection	381	0.228	0.676
Partial/half-answer	79	0.000	0.705

Table 5: Per-class accuracy and overall mean prediction confidence (Conf) from out-of-fold predictions on the training set. Note that Conf is computed across all instances (both correct and incorrect) for a given class.

even when isolating exclusively the misclassified instances (*Ambivalent*: 0.898, *Clear Reply*: 0.911, *Clear Non-Reply*: 0.888). Due to severe imbalance, *Ambivalent* (59.2% of train) acts as a majority-class sink, absorbing 35.6% of *Clear Reply* and 30.6% of *Clear Non-Reply* instances, whereas only 16.9% of *Ambivalent* cases are misclassified as *Clear Reply*. This asymmetry suggests our unweighted loss biases the model to treat *Ambivalent* as a high-probability prior for borderline cases.

Subtask 2 (Evasion) Lexical surface cues clearly differentiate performance across categories. Formulaic classes (*Clarification*: 70.7%, *Explicit*: 64.3%, *Declining to answer*: 60.7%) outperform implicature-reliant categories (*General*: 30.6%, *Implicit*: 28.1%, *Deflection*: 22.8%) and exhibit higher mean confidence (0.835 vs 0.698, frequency-weighted). Notably, *Partial/half-answer* ($n = 79$) yields zero recall, with predictions scattered across seven classes. Furthermore, *Implicit* is frequently misclassified as *Explicit* (28.1%), while *Deflection* splits between *Explicit* (20.7%) and *Dodging* (18.9%), reflecting semantic overlap among non-refusal redirections.

Model Errors and Annotator Disagreement On the dev set, model errors correlate strongly with human disagreement. Evasion majority-vote agreement averages 44.0% overall but diverges sharply by consensus: 58.4% on unanimous items ($n = 125$) versus 32.0% on 2-1 splits ($n = 150$). While any-annotator agreement reaches 59.1%, this lenient metric inflates amid maximally ambiguous 1-1-1 cases. Subtask 1 exhibits a parallel trend, with clarity accuracy dropping from 80.0% (unanimous evasion cases) to 68.7% (2-1 splits). These results indicate that a substantial fraction of model errors occurs on inherently ambiguous instances where annotator consensus is weak.

Qualitative Analysis The qualitative examples in Appendix D reveal a shared failure mode: over-reliance on surface forms, where assertive language masks evasive pragmatics. These failures highlight that simple input concatenation fails to force the model to verify if the response actually resolves the question's core demand.

6 Conclusion

We presented a hierarchical multi-task system for SemEval-2026 Task 6 (CLARITY) designed for long-context political interview responses. Our approach combines overlapping sliding-window chunking with Max-Pooling aggregation to preserve full-response evidence, and jointly trains two classification heads. Our ensemble achieved a Macro-F1 of 0.80 on Subtask 1 and 0.51 on Subtask 2, ranking 11th in both subtasks.

Max-Pooling consistently outperformed Mean-Pooling and the first-chunk baseline, confirming that evasion cues are often localized. Error analysis revealed that inherent ambiguity remains challenging: performance degrades on subtle minority categories often confused with explicit or implicit answers, whereas strategies with strong lexical cues maintain high recall.

Future work will explore targeted augmentation for minority classes. Additionally, question–answer concatenation may fail to capture the relational structure underlying evasion. We therefore plan to investigate Natural Language Inference (NLI) frameworks and dual-encoder late-interaction models such as ColBERT (Khattab and Zaharia, 2020), which allow questions to query token-level answer representations.

Acknowledgments

We thank the SemEval-2026 Task 6 organizers for the CLARITY dataset and evaluation platform, and the anonymous reviewers for their valuable feedback that improved this work.

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Peter Bull. 2003. *The Microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge, London.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14(4):651–666.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Steven E. Clayman. 2001. Answers and evasions. *Language in Society*, 30(3):403–442.

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. Association for Computational Linguistics.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.

Sandra Harris. 1991. Evasive action: How politicians respond to questions in political interviews. In Paddy Scannell, editor, *Broadcast Talk*, pages 76–99. Sage, London.

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2127–2136. PMLR.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48. Association for Computing Machinery.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Parameswary Rasiah. 2010. A framework for the systematic analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42(3):664–680.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

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.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

A Training Hyperparameters

Hyperparameter	Value
Optimizer	AdamW (weight decay 0.01)
Learning rate	1e−5 (10% warmup)
Batch size	8
Max epochs	15 (early stopping patience = 3)
Classifier dropout	0.1
Gradient clipping	max norm 1.0
Precision	BF16 mixed precision
Gradient checkpointing	enabled
Random seed	42 (base; per-fold offset)

Table 6: Training configuration and hyperparameters.

B Implementation Details and Dependencies

Our system was developed and evaluated using the following libraries and specific versions:

`torch` (≥2.2.2): https://pytorch.org
`transformers` (≥4.40.0): https://huggingface.co/docs/transformers
`datasets` (≥2.19.0): https://huggingface.co/docs/datasets
`accelerate` (≥0.30.0): https://huggingface.co/docs/accelerate
`scikit-learn` (≥1.4.2): https://scikit-learn.org
`numpy` (≥1.26.4): https://numpy.org
`pandas` (≥2.2.2): https://pandas.pydata.org
`protobuf` (==3.20.3): https://protobuf.dev
`sentencepiece` (≥0.2.0): https://github.com/google/sentencepiece

C Confusion Matrices

In this section, we provide the row-normalized confusion matrices for both subtasks to visualize the recall performance across classes.

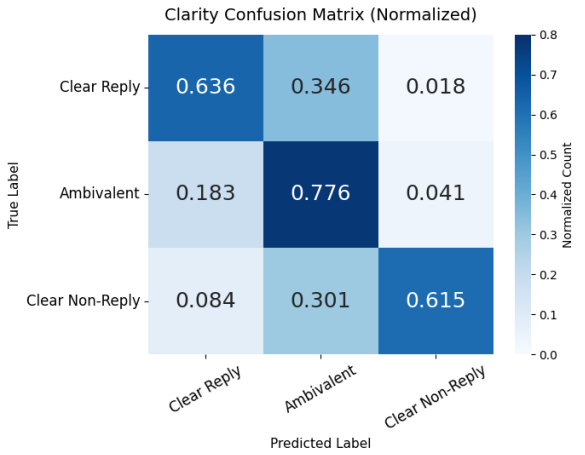

Figure 3: Normalized Confusion Matrix for Subtask 1 (Clarity)

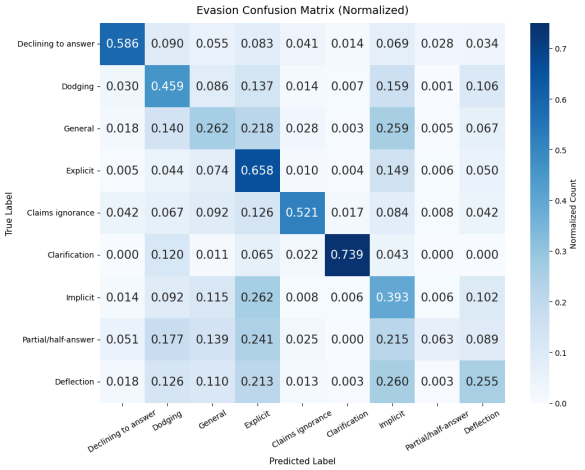

Figure 4: Normalized Confusion Matrix for Subtask 2 (Evasion)

D Qualitative Error Examples

Table 7 exemplifies the error patterns in Section 5.3, specifically the model's reliance on surface lexical cues over pragmatic inference. While explicit meta-refusals (Example 1) enable correct, high-confidence (0.990) predictions of *Clear Non-Reply* and *Declining to answer*, more subtle strategies often lead to misclassification. In Example 2, the model incorrectly predicts *Explicit* for a *Partial/half-answer*, prioritizing a direct-sounding response about timing while overlooking the omitted location. Similarly, in Example 3, assertive first-person surface features override the pragmatic intent of ironic humor, causing a *Deflection* to be misread as an *Explicit Clear Reply*. These instances reinforce that simple concatenation fails to force the model to verify if responses resolve the question's core demand.

Question (abridged)	Answer (abridged)	Gold		Predicted	
		Clar.	Eva.	Clar.	Eva.
Do you think it was the right action for Israel to take?	*I'm not going to comment on the subject that you're trying to get me to comment on.*	CNR	DTA	CNR	DTA
Your acceptance speech — are you physically going to be in Charlotte, or will you give the speech here?	*We'll be doing a speech on Thursday — the main speech. Charlotte — they will be doing nominations on Monday. I speak on Thursday.*	AMB	PHA	AMB	EXP
Do you think that if there is a breach, nobody is going to blame you?	*Of course, no one would blame me. I know you won't. You'll be saying, Biden did a wonderful job.*	AMB	DEF	CR	EXP

Table 7: Representative prediction outcomes from out-of-fold analysis. Abbreviations—Clar.: Clarity; Eva.: Evasion; CNR: Clear Non-Reply; CR: Clear Reply; AMB: Ambivalent; DTA: Declining to answer; EXP: Explicit; PHA: Partial/half-answer; DEF: Deflection.