AI@UMS at SemEval-2026 Task 6: Handling Long Question-Answer Pairs with Sliding-Window Models for Clarity and Evasion Analysis

Ikhlasul Amal[1], Zia Ul Rehman Zafar[2], Choiru Yahya Firdaus[2], Endang Wahyu Pamungkas[2]
[1]Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
ikhlasulamal@mail.ugm.ac.id,
{l200249293,l200234172}@student.ums.ac.id,
ewp123@ums.ac.id

Abstract

This paper presents the AI@UMS system for SemEval-2026 Task 6: CLARITY – Unmasking Political Question Evasions. The task requires classifying question-answer (QA) pairs from political interviews along two dimensions: clarity level (Subtask 1) and evasion technique (Subtask 2). A key challenge is that political interview transcripts often exceed the 512-token input limit of standard transformer encoder models. We address this with a sliding-window fine-tuning strategy applied to `roberta-base`, where each QA pair is segmented into overlapping windows of 512 tokens with a stride of 256 tokens. Per-window predictions are aggregated via softmax probability averaging across multiple windows and across an ensemble of three independently trained models with different random seeds. We further employ class-weighted focal-inspired loss and label smoothing to mitigate the pronounced class imbalance in both subtasks. Our system achieves macro F1 scores of 0.62 (Subtask 1, rank 34/40) and 0.48 (Subtask 2, rank 14/33) on the official evaluation set.

1 Introduction

Political evasion and equivocation are well-documented phenomena in discourse analysis. Bull (2003) reports that politicians provided clear answers to only 39–46% of questions in televised interviews, compared to 70–89% for non-politicians, illustrating the strategic ambiguity inherent to political speech. Automating the detection and classification of such evasion strategies has attracted growing interest from both the NLP and political science communities (Thomas et al., 2024; Ferracane et al., 2021).

SemEval-2026 Task 6, CLARITY (Thomas et al., 2026), formalises this problem as two classification tasks over QA pairs extracted from US presidential interviews. Subtask 1 requires assigning one of three *clarity-level* labels: CLEAR REPLY, AMBIVALENT REPLY, or CLEAR NON-REPLY. Subtask 2 requires assigning one of nine fine-grained *evasion technique* labels derived from political discourse typologies (Bull and Strawson, 2019; Rasiah, 2010).

The task builds on the dataset and taxonomy introduced by Thomas et al. (2024), who demonstrated that encoder models such as RoBERTa (Zhuang et al., 2021) and DeBERTa (He et al., 2021) struggle when applied naively to long QA pairs: only 63% of training samples fit within the 512-token limit of standard encoders, meaning a significant portion of context is lost through truncation. This finding motivates our central design choice: a sliding-window segmentation strategy (Pappagari et al., 2019; Li et al., 2022) that allows the model to process arbitrarily long inputs without discarding information.

Transformer-based pre-trained language models have demonstrated strong performance across a wide range of NLP tasks (Devlin et al., 2019; Zhuang et al., 2021; He et al., 2021). In the context of long document classification, several works have proposed chunking or segmentation strategies to bypass the fixed-length input constraint (Pappagari et al., 2019; Beltagy et al., 2020; Dai et al., 2022). Our approach draws on this line of work, adapting a simple sliding-window mechanism to the sequence classification setting of the CLARITY task.

To further address the class imbalance that characterises both subtasks – with EXPLICIT responses dominating the evasion distribution – we employ class-weighted loss combined with a focal-loss-inspired modulating factor (Lin et al., 2020), along with label smoothing (Szegedy et al., 2016). We also apply a multi-seed ensemble (Lakshminarayanan et al., 2017) to improve prediction stability.

Our main contribution is adapting a sliding-window segmentation strategy for RoBERTa to handle the long QA pairs characteristic of polit-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3544–3548

ical interview data, combined with class-weighted focal-inspired loss, label smoothing, and multi-seed ensembling to address class imbalance and training variance.

2 Background

2.1 Task Description

The CLARITY shared task (Thomas et al., 2026) operates on the QEvasion dataset (Thomas et al., 2024), which comprises 3,445 question-answer pairs drawn from interviews of four US presidents. The dataset is annotated following a two-level hierarchical taxonomy: (1) a high-level *clarity level* (Clear Reply, Ambivalent Reply, Clear Non-Reply) and (2) a fine-grained *evasion technique* level with nine subcategories (Explicit, Implicit, General, Partial, Dodging, Deflection, Declining to answer, Claims ignorance, Clarification). Both subtasks are evaluated using macro F1-score.

2.2 The Long-Context Problem for Encoders

Pre-trained encoder models such as BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021) operate with a fixed maximum input length of 512 tokens. For tasks involving long documents or transcripts, this limit causes information loss when inputs are naively truncated. Several strategies have been proposed to mitigate this: (i) truncation to the first 512 tokens (Li et al., 2022), (ii) sparse attention models such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020), and (iii) chunking or sliding-window approaches that segment the input and aggregate segment-level predictions (Pappagari et al., 2019; Dai et al., 2022).

Thomas et al. (2024) explicitly identify truncation as a major source of error for encoder models in the CLARITY setting: only 63% of training examples fit within 512 tokens, and performance on the remaining 37% is close to random chance. This strongly motivates a sliding-window approach for this task.

2.3 Imbalanced Classification in NLP

Both subtasks exhibit significant class imbalance: in the evasion-level distribution, EXPLICIT replies account for roughly 30% of samples, while minority classes such as PARTIAL/HALF-ANSWER and CLARIFICATION each comprise fewer than 3% of examples (Thomas et al., 2024). Standard cross-entropy loss tends to bias predictions toward majority classes in such settings. Focal loss (Lin et al., 2020), originally proposed for object detection, addresses this by down-weighting easy examples via a modulating factor $(1 - p_t)^\gamma$, encouraging the model to focus on harder, minority-class samples. Class-weighted loss is a complementary strategy that assigns higher penalties to misclassifications of underrepresented classes (King and Zeng, 2001).

3 System Description

3.1 Model Architecture

We use `roberta-base` (Zhuang et al., 2021) as our encoder backbone, a robustly optimised variant of BERT (Devlin et al., 2019) pre-trained on 160GB of English text. On top of the encoder, we add a dropout layer (rate = 0.3) and a single linear classification head mapping from the hidden dimension (768) to the number of target labels (3 for Subtask 1, 9 for Subtask 2). Classification uses the `[CLS]` token representation from the final transformer layer.

We augment the tokenizer with four domain-specific special tokens: `<question>`, `</question>`, `<answer>`, `</answer>`, to explicitly delimit the question and answer spans within each input.

3.2 Sliding-Window Segmentation

To address the long-context limitation of `roberta-base`, we implement a sliding-window strategy at the token level. Given a tokenised input sequence of length L, we extract consecutive overlapping windows of size $W = 512$ tokens with stride $S = 256$ tokens, yielding an overlap of 256 tokens between consecutive windows. This 50% overlap ensures that contextual information near window boundaries is captured by at least two windows, mitigating the context fragmentation problem identified in prior work (Dai et al., 2022; Pappagari et al., 2019). The final window is zero-padded to W if necessary.

During training, each window is treated as an independent sample with the label of its parent example. During inference, all windows belonging to the same example are passed through the model independently, producing per-window logit vectors. These are converted to probability distributions via softmax, and the final prediction is obtained by averaging the per-window probabilities and taking the argmax of the resulting distribution. Figure 1 provides a schematic overview of the pipeline.

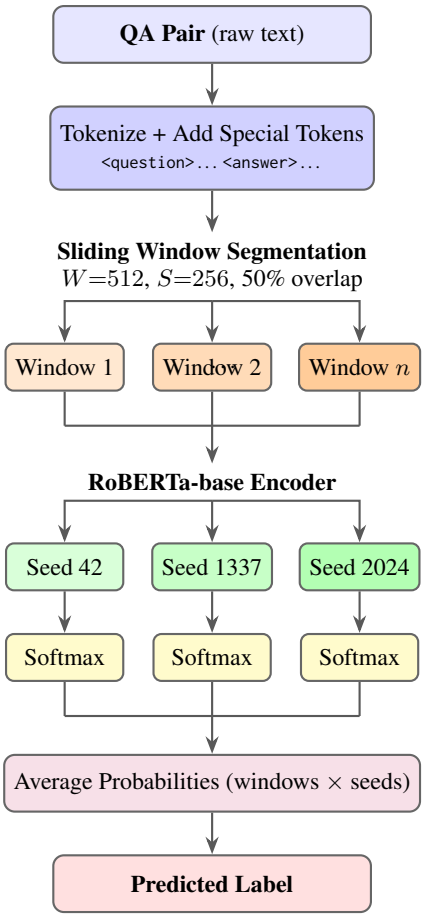

Figure 1: Overview of the AI@UMS pipeline. A QA pair is tokenised with domain-specific special tokens and segmented into overlapping windows (W=512, S=256). Each window is independently encoded by `roberta-base`. Three models trained with different random seeds produce per-window softmax probabilities, which are averaged across all windows and seeds to yield the final predicted label.

3.3 Loss Function and Class Imbalance Handling

We address class imbalance through two complementary mechanisms. First, we compute class weights using the balanced weighting strategy from scikit-learn, where the weight of class c is inversely proportional to its frequency. These weights are incorporated into the cross-entropy loss as per-class scaling factors. Second, we apply a focal-loss-inspired modulation factor (Lin et al., 2020), scaling each sample's loss by $(1 - p_t)^\gamma$ where p_t is the predicted probability for the ground-truth class and $\gamma = 1.0$. This combination penalises confident correct predictions less, focusing training capacity on hard and minority-class examples. We additionally apply label smoothing (Szegedy et al., 2016) with $\epsilon = 0.05$ to reduce overconfidence.

3.4 Training Details

We train the model for 12 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate 3×10^{-5} and weight decay 0.01. Learning rate is scheduled with a cosine decay following a linear warmup over the first 10% of training steps (Loshchilov and Hutter, 2017). We use gradient accumulation over 4 steps with an effective batch size of 16, and mixed-precision training (FP16) via PyTorch AMP. All experiments are conducted on a single NVIDIA Tesla T4 GPU (Google Colab).

3.5 Multi-Seed Ensemble

To reduce prediction variance arising from random initialisation and stochastic training, we train three independent models with different random seeds (42, 1337, 2024) (Lakshminarayanan et al., 2017). During inference, we collect the per-window softmax probabilities from all three models and average them before taking the argmax. This seed ensemble strategy is computationally inexpensive relative to ensembling different model architectures and consistently improves stability in low-resource classification settings.

Both Subtask 1 and Subtask 2 use the same architecture and training procedure; the only difference is the output dimensionality of the classification head (3 vs. 9 classes) and the corresponding label mapping.

4 Results

Table 1 presents the official macro F1 scores for our system alongside the top-performing system. Our system achieves macro F1 = 0.62 on Subtask 1 and 0.48 on Subtask 2, while the top-performing TeleAI system achieves 0.89 and 0.68 respectively. The score on Subtask 1 reflects the benefit of processing full QA pairs without truncation: political interview answers frequently contain discourse-level cues spread across the entire response, and the sliding-window strategy with 50% overlap allows the model to observe these cues rather than losing them at the 512-token boundary (Dai et al., 2022; Thomas et al., 2024). The multi-seed ensemble further contributes to the robustness of this score by smoothing over the variance that arises from individual training runs on a relatively small dataset of approximately 2,700 examples (Lakshminarayanan et al., 2017).

The lower score on Subtask 2 (0.48) reflects the

Subtask	Team	Rank	Macro F1
Subtask 1	TeleAI	1/40	0.89
	AI@UMS	**34/40**	**0.62**
Subtask 2	TeleAI	1/33	0.68
	AI@UMS	**14/33**	**0.48**

Table 1: Official leaderboard results for both subtasks.

considerably harder nature of nine-class evasion classification, where subtle distinctions between categories such as DODGING, DEFLECTION, and GENERAL are challenging even for human annotators (Thomas et al., 2024). Despite applying class-weighted loss and focal-inspired modulation to counteract the skewed label distribution – where EXPLICIT replies dominate at roughly 30% while categories like CLARIFICATION fall below 3% – the model still struggles to consistently identify the rarest evasion techniques (Lin et al., 2020; King and Zeng, 2001). This suggests that the training signal available for minority classes remains insufficient, and that the window-level predictions are less reliable when the evasion cue is subtle and localised to a small portion of the answer.

Overall, the performance gap with respect to the leading system points to the limits of a compact encoder-based approach under constrained computational resources. Systems leveraging larger language models or architectures with native long-context support (Beltagy et al., 2020; Zaheer et al., 2020) are likely better positioned to capture the nuanced reasoning required for both subtasks. Nevertheless, our results demonstrate that a carefully designed sliding-window pipeline with ensemble aggregation can serve as a strong and resource-efficient baseline for this task.

5 Conclusion

We presented the AI@UMS system for SemEval-2026 Task 6 (CLARITY), which adapts `roberta-base` to handle long political QA pairs through a sliding-window segmentation strategy, multi-seed ensembling, and class-imbalance-aware loss. Our system achieves macro F1 = 0.62 on Subtask 1 and 0.48 on Subtask 2, outperforming the truncation-based baseline for Subtask 1 and placing in the top half of participants for Subtask 2. Future work will explore hierarchical aggregation of window representations, data augmentation for minority evasion classes, and larger pre-trained models with extended context windows.

Limitations

Our approach has several limitations. First, the sliding-window strategy treats windows as independent inputs during training, which does not model inter-window dependencies that may carry discourse-level cues relevant to evasion classification. Second, our experiments are constrained to `roberta-base` due to computational resources (single T4 GPU on Google Colab), and larger or domain-adapted models may perform substantially better. Third, the class imbalance in the evasion subtask remains a challenge, as minority classes are underrepresented in the training data even after class-weighted loss is applied.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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

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. volume 73, pages 429–449.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT with DISENTANGLED AT-

TENTION. In *International Conference on Learning Representations*.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9:137–163.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6405–6416. Curran Associates Inc.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. volume 42, pages 318–327.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

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. Curran Associates Inc.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Hyperparameter Configuration

Hyperparameter	**Value**
Base model	`roberta-base`
Window size (W)	512 tokens
Stride (S)	256 tokens
Batch size	4
Gradient accumulation	4 steps
Effective batch size	16
Learning rate	3×10^{-5}
Weight decay	0.01
Warmup ratio	0.10
LR schedule	Cosine decay
Dropout	0.3
Focal γ	1.0
Label smoothing ϵ	0.05
Epochs	12
Ensemble seeds	42, 1337, 2024
Precision	FP16 (AMP)
Hardware	NVIDIA Tesla T4

Table 2: Hyperparameter configuration used for both subtasks.