B&B at SemEval-2026 Task 6: A RoBERTa-based Model with NLI-derived Semantic Features for Clarity-Level Classification of Political Question Evasion

Chi-Bo Lin
University of Colorado Boulder
Bob.Lin@colorado.edu

Boyang Yu
University of Colorado Boulder
boyu5591@colorado.edu

Abstract

Equivocation and ambiguity are common phenomena in political interviews, where public figures often avoid providing direct answers to challenging questions. We present our submission to SemEval-2026 Task 6, Subtask 1 on political response clarity classification. Our system builds on RoBERTa and incorporates NLI-derived semantic features to better distinguish *Clear Reply*, *Ambivalent*, and *Clear Non-Reply* responses. To address class imbalance and performance instability, we explore class weighting, multi-seed ensembling, and a hierarchical two-stage framework with threshold tuning. Our best model achieves 60% macro-F1 on the official test set and 64% macro-F1 on an additional evaluation set, demonstrating stable performance across splits. While large language models have shown strong zero-shot performance on related discourse tasks, our results indicate that carefully engineered smaller models, when combined with structured semantic features and imbalance-aware training, provide an effective and computationally efficient solution for political response clarity classification under limited training data.

1 Introduction

The CLARITY – Unmasking Political Question Evasions task (SemEval-2026 Task 6) (Thomas et al., 2024, 2026) aims to automatically detect the clarity and evasiveness of political responses in English interview transcripts. Political figures often avoid providing direct answers, making automated clarity analysis important for political transparency, media accountability, and computational social science research. The task provides a ChatGPT-assisted and human-verified dataset of U.S. political interviews and challenges systems to determine how clearly an answer addresses its corresponding question. We focus on Subtask 1: Clarity-Level Classification, a three-way single-label task distinguishing between *Clear Reply*, *Ambivalent*, and *Clear Non-Reply* responses.

Our system builds upon a RoBERTa encoder enhanced with external semantic signals. Specifically, we incorporate Natural Language Inference (NLI) probabilities—*entailment*, *neutral*, and *contradiction*—to better capture the semantic relationship between each question–answer pair. To address label imbalance in the training data and reduce variance under limited supervision, we apply inverse-frequency class weighting, multi-seed ensemble, and a hierarchical two-stage classification framework. We further conduct systematic hyperparameter tuning and threshold optimization to improve macro-F1 performance and balance predictions across classes.

Our experiments show that incorporating NLI-derived semantic features improves macro-F1 by providing additional semantic signals about the relationship between questions and answers. Imbalance-aware training strategies and hierarchical modeling further stabilize performance across classes. After optimization, our best model achieves 60% macro-F1 on the official test set (308 instances) and 64% macro-F1 on an additional evaluation set (237 instances), demonstrating consistent performance across splits. To facilitate future research, we release all code and configuration files under an open-source license.[1]

2 Background

2.1 Task Setup

Subtask 1 of SemEval-2026 Task 6 (CLARITY) requires classifying the clarity of a political response given a question–answer pair. The system predicts one of three labels: *Clear Reply*, where the answer directly addresses the question and admits only a single interpretation; *Clear Non-Reply*, where the speaker avoids answering or explicitly refuses to provide information; and *Ambivalent*,

[1] https://github.com/BobLinChiBo/Clarity_SemEval_2026_BB

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3060–3067

where the response is partially relevant but permits multiple interpretations.We use the QEvasion dataset—English political interview transcripts collected from the official White House website and spanning four U.S. presidents: George W. Bush, Barack Obama, Donald J. Trump, and Joseph R. Biden, from 2006 to 2023. These transcripts are transformed into question–answer pairs, resulting in 3,448 labeled training instances, 308 labeled test instances, and 237 additional evaluation instances (Thomas et al., 2024). Each instance consists of an interviewer's question, the interviewee's answer, and an associated clarity label. Performance is evaluated using macro F1-score, which is particularly appropriate given the substantial class imbalance in the training set. The detailed class distribution of the training data is shown in Appendix Table 4.

2.2 Related Work

Determining whether a political response meaningfully addresses a question requires modeling semantic alignment between the question and answer. Natural Language Inference (NLI) offers a principled framework for capturing such relationships through entailment, contradiction, and neutrality. Prior work shows that NLI-trained models such as RoBERTa–MNLI learn relational semantic representations that transfer effectively to downstream reasoning and alignment tasks (Liu et al., 2019; Poliak et al., 2018; Li et al., 2019). Because political interview responses are often long and context-rich, we additionally draw on work emphasizing the importance of contextual reasoning for NLI over long-form text (Liu et al., 2020). Moreover, domain-adapted NLI models have been shown to serve as efficient and reproducible classifiers for political text, sometimes outperforming large generative LLMs in zero- and few-shot political classification settings (Burnham et al., 2025).These findings motivate our incorporation of NLI probabilities as auxiliary features for the political question–answer clarity task. NLI outputs provide interpretable signals that correlate with answering behavior; for example, a high contradiction probability often aligns with evasive or non-reply responses, whereas entailment is more commonly associated with clear replies.

Political question–answer clarity classification faces two key practical challenges: severe class imbalance and training instability. In real-world interview data, ambivalent responses substantially outnumber clear replies and explicit non-replies, causing standard flat multi-class cross-entropy training to bias optimization toward majority classes and often reducing minority-class recall (Buda et al., 2018; Henning et al., 2023). While prior work often relies primarily on loss reweighting, less attention has been given to architectural approaches that reduce direct competition among highly imbalanced labels. To address this gap, we combine inverse-frequency class weighting with a hierarchical classification framework that first separates replies from non-replies before refining minority distinctions, improving sensitivity to rare categories such as *Clear Non-Reply*. Additionally, limited dataset size introduces substantial variance across random initializations, with validation macro-F1 varying from 64–70% in our experiments. Following prior work on seed sensitivity and ensemble robustness (Bui et al., 2025; Lakshminarayanan et al., 2017), we mitigate this instability through multi-seed ensembling, averaging predictions across independently initialized models to improve reproducibility and generalization.

Taken together, our approach integrates NLI-derived semantic signals with contextual representations from RoBERTa while addressing class imbalance and training instability in political response clarity classification.

3 System Overview

To study the integration of NLI signals with RoBERTa, we first establish a baseline model that directly concatenates the RoBERTa `[CLS]` representation with the raw three-dimensional NLI probability vector, without any gating mechanism. Because NLI signals may be noisy or imperfectly aligned with the clarity labels, we extend this baseline by introducing a lightweight NLI adapter and a CLS-driven scalar gate. Specifically, the NLI probabilities are projected through a small MLP and scaled by a learned gate before being concatenated with the RoBERTa CLS representation for classification (Figure 1).

The boundary between *Ambivalent* and the other two categories (*Clear Reply* and *Clear Non-Reply*) is often subjective and difficult to learn consistently. To address this challenge, we further explore a hierarchical two-stage classification framework. In Stage 1, the model distinguishes between *Clear Reply* and *Non-Reply-like* responses (i.e., *Ambivalent* and *Clear Non-Reply*). In Stage 2, applied

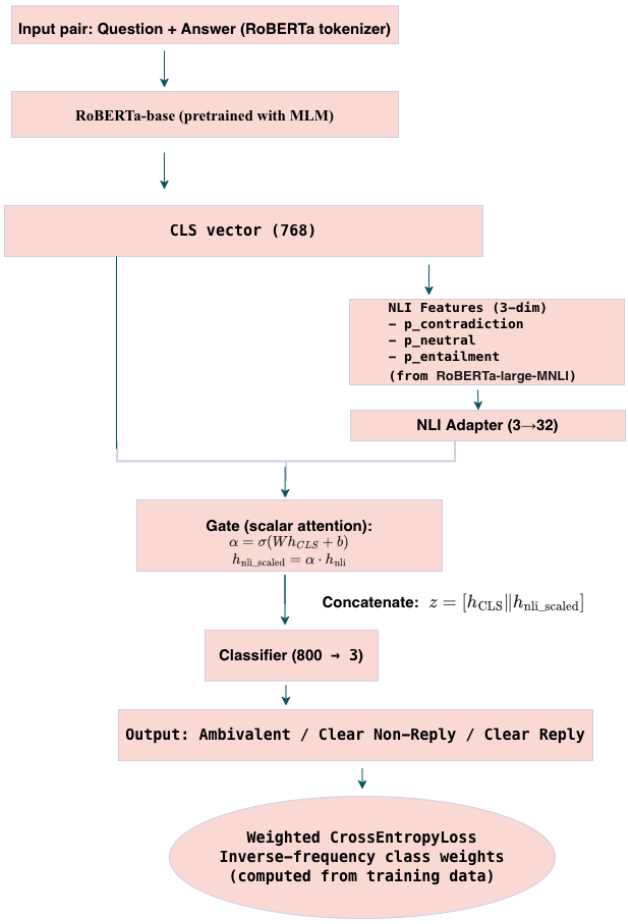

Figure 1: Overview of the gated NLI-enhanced RoBERTa architecture

only when Stage 1 predicts the latter group, the model further differentiates between *Ambivalent* and *Clear Non-Reply*. A schematic overview of the hierarchical system is shown in Figure 2.

3.1 Precomputed NLI Features

To capture the semantic relation between each interview question and answer, we precompute Natural Language Inference (NLI) features using RoBERTa-large-MNLI, a RoBERTa model fine-tuned on MNLI (Liu et al., 2019; Wang et al., 2019). For each (question, answer) pair, we treat the answer as the premise and the question as the hypothesis, and extract the model's posterior probabilities over *contradiction*, *neutral*, and *entailment*, i.e., $(p_{\text{contra}}, p_{\text{neutral}}, p_{\text{entail}})$. While MNLI labels are not strict logical judgments for interrogative hypotheses, these scores provide an interpretable proxy for semantic alignment: higher entailment often corresponds to direct replies, contradiction can indicate semantic mismatch or opposition, while neutrality often captures partial, indirect, or off-topic responses (Mitra et al., 2020). We compute these features once and store them alongside the dataset, enabling downstream training to incorporate NLI semantics without running NLI inference during each epoch.

3.2 Gated Integration

In addition to direct concatenation of NLI features with the RoBERTa representation, we also explore a CLS-driven scalar gating mechanism that adaptively controls the contribution of NLI signals. RoBERTa encodes the question–answer pair and produces a contextual embedding h_{CLS} from the first token representation. The NLI probability vector $p_{NLI} \in \mathbb{R}^3$ is projected through a multilayer perceptron (MLP) to obtain a dense representation:

$$h_{\text{nli}} = \text{MLP}\left(p_{\text{NLI}}\right)$$

A scalar gate is computed from the CLS representation:

$$\alpha = \sigma\left(W h_{\text{CLS}} + b\right)$$

where $W \in \mathbb{R}^{1\times 768}$. The NLI representation is then scaled:

$$h_{\text{nli\_scaled}} = \alpha \cdot h_{\text{nli}}$$

Finally, the scaled NLI representation is concatenated with the CLS embedding:

$$z = [h_{\text{CLS}} \| h_{\text{nli\_scaled}}]$$

which is passed to a three-way classifier for clarity prediction.

3.3 Class Weighting for Label Imbalance

The QEvasion dataset exhibits moderate class imbalance, with *Ambivalent* as the majority class. We therefore apply inverse-frequency class weighting in the cross-entropy loss:

$$w_c = \frac{N}{K \cdot n_c},$$

where N is the number of training examples, $K = 3$ the number of classes, and n_c the frequency of class c. This increases the contribution of minority classes such as *Clear Non-Reply*.

To reduce variance due to random initialization, we employ a multi-seed ensemble. We train three independent instances of the model with different random seeds (21, 42, and 84), each for 10 epochs. During inference, predictions are aggregated by averaging the output logits from the three models

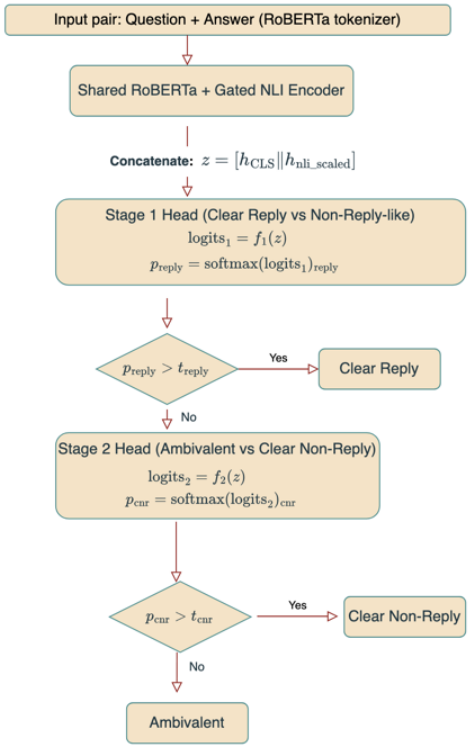

Figure 2: Overview of the hierarchical two-stage classification framework

and selecting the class with the highest mean logit. This logit-level ensembling reduces sensitivity to stochastic optimization and improves prediction stability.

3.4 Hierarchical Two-stage Classification Framework

Although class weighting and multi-seed ensembling improve training stability, the model still exhibits uneven performance across classes, particularly between the majority class (*Ambivalent*) and the minority classes (*Clear Reply* and *Clear Non-Reply*). To better balance prediction errors, we introduce a hierarchical two-stage classification framework that decomposes the three-way prediction into sequential binary decisions.

In Stage 1, the model distinguishes *Clear Reply* from *Non-Reply-like* responses (i.e., *Ambivalent* and *Clear Non-Reply*). In Stage 2, applied only when Stage 1 predicts *Non-Reply-like*, the model further differentiates *Ambivalent* from *Clear Non-Reply* (Figure 2).

3.5 Hyperparameter and Threshold Tuning

We conduct a two-phase validation-based random search over model and training hyperparameters, including learning rate, head learning-rate multiplier, warmup ratio, weight decay, gradient clipping, and number of epochs. In Phase 1, configurations are sampled broadly from predefined ranges; in Phase 2, sampling is refined around the best-performing validation configuration. For each trial, stage-specific inference thresholds t_{reply} and t_{cnr} are tuned on the same validation split used for model selection. Thresholds are first searched using a coarse grid from 0.50 to 0.95 in increments of 0.05, followed by fine-grained refinement within ± 0.10 of the best coarse thresholds using step size 0.01. Final model selection is based on validation macro-F1.

4 Experimental Setup

The QEvasion dataset contains 3,448 labeled training instances, 308 test instances, and 237 additional evaluation instances provided by the task organizers to verify robustness across different splits. Following the shared task guidelines, we split the original training portion into 90% for training and 10% for validation. All experiments are conducted on this split unless otherwise noted.

We begin with a RoBERTa-base baseline model. For each question–answer pair, we precompute natural language inference (NLI) features using the pretrained `roberta-large-mnli` model[2]. This step computes three semantic relationship probabilities—*contradiction*, *neutral*, and *entailment*—which are stored as continuous features and reused during training, thereby avoiding repeated NLI inference. We integrate these NLI features into the classifier both with and without a gating mechanism.

To mitigate class imbalance, we apply inverse-frequency class weighting during cross-entropy optimization and further explore a hierarchical two-stage classification framework. During preliminary experiments, we observe noticeable performance variance across random initializations (64%–70% macro-F1 on the validation set), motivating the use of multi-seed ensembling to improve stability. We additionally conduct hyperparameter tuning and select model configurations based on validation

[2]Model available at: https://huggingface.co/roberta-large-mnli

macro-F1, following the procedure described in Section 3.5.

Following Thomas et al. (2024), we report macro-F1 as the primary evaluation metric:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{c=1}^{K} F1_c,$$

where K denotes the number of classes.

5 Results

5.1 Ablation Study

Table 1 presents an ablation-style comparison of our primary model variants. Starting from a RoBERTa-base baseline, we incrementally incorporate NLI features, class weighting, gated fusion, multi-seed ensembling, and hierarchical decoding to evaluate the contribution of each component. The best-performing configuration (highlighted in red in Table 1) is used for the final Subtask 1 submission.

Incorporating NLI features yields the largest single-stage improvement over the RoBERTa-base baseline, increasing macro-F1 substantially on both test and evaluation sets. This suggests that semantic alignment signals between interview questions and responses provide highly informative auxiliary features for political clarity classification. Directly modeling contradiction, neutrality, and entailment probabilities improves the model's ability to distinguish among *Ambivalent*, *Clear Reply*, and *Clear Non-Reply* categories.

Gated fusion provides modest additional gains by adaptively controlling the influence of NLI-derived semantic signals based on contextual encoder representations, reducing overreliance on potentially noisy external probabilities. Multi-seed ensembling improves robustness and test-set performance by reducing variance across random initializations, though evaluation-set gains remain limited.

The hierarchical two-stage framework provides the strongest overall performance. By decomposing the original three-way classification problem into sequential binary decisions, the model simplifies difficult semantic distinctions and allows stage-specific weighting and threshold tuning. Comparing the hierarchical two-stage models with single-stage baselines, we observe consistent improvements in macro-F1 on both the test and evaluation sets. The hierarchical formulation improves performance on *Ambivalent* and *Clear Reply* (Appendix Table 6 and 7), while maintaining competitive performance on *Clear Non-Reply*. Overall, decomposing the original three-way task into simpler binary decisions helps the model better balance classification errors across labels and leads to more stable performance across validation and test splits. The final model configuration selected via validation macro-F1 is detailed in Appendix Table 5.

5.2 Error Analysis

Despite the substantial gains from NLI integration and hierarchical modeling, several persistent challenges remain.

Applying naive inverse-frequency class weighting does not consistently improve overall performance. On the evaluation dataset, while *Clear Reply* F1 improves substantially (0.51 $\rightarrow$ 0.59), both *Ambivalent* and *Clear Non-Reply* decline, reducing macro-F1 from 0.63 to 0.59. A similar pattern appears on the validation set, where weighting modestly improves some minority-class performance but degrades broader classification balance (Tables 2 and 3). These findings suggest that standard inverse-frequency reweighting introduces unstable class trade-offs rather than uniformly improving minority-class sensitivity.

A central limitation of single-stage classification is that it forces semantically overlapping categories—particularly *Ambivalent* and *Clear Non-Reply*—to compete within a single decision space. Small probability shifts can therefore produce unstable class assignments. The hierarchical framework alleviates this issue by isolating reply detection from finer-grained non-reply distinctions, improving both interpretability and class-specific robustness.

Overall, our findings suggest that while semantic augmentation substantially improves political response clarity classification, effective performance ultimately depends on combining semantic features with architectures specifically designed to manage class imbalance, semantic overlap, and optimization instability.

6 Conclusion and Limitations

Our results demonstrate that integrating NLI-derived semantic signals improves clarity classification, particularly under class imbalance. While inverse-frequency reweighting alone yields inconsistent gains, hierarchical modeling provides a more stable and structured solution by decompos-

Model	Macro-F1 (Test)	Macro-F1 (Eval)	Description
RoBERTa-base	45%	59%	No NLI features
+ NLI (concat)	53%	63%	Direct concatenation
+ class weighting	50%	59%	Inverse-frequency loss
+ gating	51%	60%	CLS-driven scalar gate
+ multi-seed ensemble	55%	58%	Logit-level ensembling (21, 42, 84)
Hierarchical Two-Stage Models			
Hierarchical (gated + stage-weighted)	59%	64%	Tuned thresholds; default hyperparameters
Hierarchical + HP search	60%	64%	Tuned thresholds; random-search hyperparameters

Table 1: Ablation-style comparison of baseline and NLI-enhanced model variants on the QEvasion test and evaluation sets

Class	F1 (No Weighting)	F1 (+Weighting)
Ambivalent	0.61	0.52
Clear Non-Reply	0.77	0.66
Clear Reply	0.51	0.59
Macro-F1	0.63	0.59

Table 2: Per-class F1 comparison with and without inverse-frequency class weighting on the evaluation dataset

Class	F1 (No Weighting)	F1 (+Weighting)
Ambivalent	0.76	0.67
Clear Non-Reply	0.67	0.69
Clear Reply	0.56	0.59
Macro-F1	0.66	0.64

Table 3: Per-class F1 comparison with and without inverse-frequency class weighting on the validation set

ing the decision process. The final model achieves 60% macro-F1 on the official test set and 64% on the evaluation set, indicating consistent generalization across data splits.

Despite these improvements, our results remain below top-performing systems in the shared task. This gap is consistent with broader findings that large language model (LLM)-based approaches substantially outperform encoder-only architectures on discourse-level reasoning tasks. As shown in the task overview (Thomas et al., 2026), top-performing systems leverage multi-stage reasoning, hierarchical decomposition, and prompting strategies, whereas encoder-based models typically require substantially more complex architectures—such as multi-task learning, graph neural networks, or multiple instance learning—to achieve competitive performance. This highlights both the efficiency and strong reasoning capabilities of LLM-based approaches.

Several limitations remain. First, the *Clear Reply* category is difficult to distinguish from *Ambivalent*, reflecting inherent ambiguity in political discourse and moderate annotator agreement reported in the dataset. Second, our approach relies on NLI signals from a general-domain MNLI model without domain adaptation, which may limit alignment with political language. Third, the relatively small dataset increases sensitivity to initialization and optimization noise, particularly for more complex models.

Future work could explore domain-adaptive NLI representations, larger or hybrid models, and tighter integration with hierarchical or multi-stage reasoning frameworks to improve robustness and performance.

References

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Nghia Bui, Guergana Savova, and Lijing Wang. 2025. Assessing the macro and micro effects of random seeds on fine-tuning large language models.

Michael Burnham, Kayla Kahn, Ryan Yang Wang, and Rachel X. Peng. 2025. Political debate: Efficient zero-shot and few-shot classifiers for political text. *Political Analysis*, page 1–15.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles.

Tianda Li, Xiaodan Zhu, Quan Liu, Qian Chen, Zhigang Chen, and Si Wei. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *CoRR*, abs/1904.12104.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2020. Natural language inference in context - investigating contextual reasoning over long texts. *CoRR*, abs/2011.04864.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Arindam Mitra, Ishan Shrivastava, and Chitta Baral. 2020. Enhancing natural language inference using new and expanded training data sets and new learning models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8504–8511.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *CoRR*, abs/1805.01042.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

A Appendix

Class	Proportion
Ambivalent	59.2%
Clear Non-Reply	10.3%
Clear Reply	30.5%
Total Instances	3,448

Table 4: Class distribution in the QEvasion training dataset

Component	Value
Model Architecture	
Backbone	RoBERTa-base
NLI encoder	roberta-large-mnli (precomputed probabilities)
NLI feature dim.	3 (contradiction / neutral / entailment)
Fusion	CLS-conditioned scalar gating + feature concatenation
Heads	Hierarchical binary (2-way $\rightarrow$ 2-way)
Dropout	0.04
Loss	$\mathcal{L} = \mathcal{L}_{\text{stage1}} + \lambda_2 \mathcal{L}_{\text{stage2}}$
λ_2	0.84
Training Hyperparameters (selected via random search)	
Epochs	8 (early stopping, patience=2)
Batch size (train / eval)	16 / 32
Max length	256
Learning rate (encoder)	3.0×10^{-5}
Head LR multiplier	3.0
Warmup ratio	0.07
Weight decay	0.0
Gradient clipping	None
Decoding / Threshold Tuning	
Threshold selection	tuned on validation set
t_{reply}, t_{cnr}	0.75, 0.55

Table 5: Final configuration of the hierarchical RoBERTa+NLI model selected by validation macro-F1

Class	Single-stage	Hierarchical + HP search
Ambivalent	0.61	0.70
Clear Non-Reply	0.77	0.66
Clear Reply	0.51	0.57
Macro-F1	0.63	0.64

Table 6: Per-class F1 comparison between the best single-stage model and the hierarchical two-stage model on the evaluation set

Class	Single-stage	Hierarchical + HP search
Ambivalent	0.64	0.65
Clear Non-Reply	0.44	0.63
Clear Reply	0.50	0.53
Macro-F1	0.53	0.60

Table 7: Per-class F1 comparison between the best single-stage model and the hierarchical two-stage model on the testing set