PFW at SemEval-2026 Task 6: Multi-Seed DeBERTa Ensembles for Political Response Clarity and Evasion Classification

Taleef Tamsal* **and Jonathan Rusert**
Purdue University Fort Wayne
{tamst01, jrusert}@pfw.edu

Abstract

This paper describes the PFW system for SemEval-2026 Task 6 (CLARITY), which addresses the classification of response clarity and evasion techniques in political interview question-answer pairs. Rather than relying on large language model prompting, we pursue a competitive non-LLM approach based on fine-tuning DeBERTa-xlarge and DeBERTa-v3-large with a multi-seed ensemble strategy: 5-fold cross-validation with 10 random seeds yields 50 models per architecture, combined through simple logit averaging. Our system achieves a macro F1 of 0.76 on Subtask 1 (clarity-level classification) and 0.50 on Subtask 2 (evasion-type classification). We also find that three post-hoc optimization techniques—learned ensemble weights, threshold calibration, and hierarchical masking—each improve out-of-fold performance yet *degrade* evaluation scores by 0.02–0.10 F1. This pattern should be interpreted cautiously: the 237-sample evaluation set likely contributes substantial variance, and two of the three degradations fall within the $\pm$0.06 95% CI expected from sampling noise. Still, the consistent directional pattern across all three prediction-level interventions provides suggestive evidence for an optimization paradox, highlighting the risk of overfitting to cross-validation predictions when evaluation data is limited. Our code is publicly available at https://github.com/Taleef7/semeval-2026-task6.

1 Introduction

Political question evasion, the practice of providing answers that deflect, dodge, or otherwise fail to address the substance of a question, is a pervasive phenomenon in political discourse and has been studied extensively in political science and communication research (Bull, 2003; Clayman, 2001; Harris, 1991). Automatically detecting and classifying such evasion is important for media accountability and computational political science, yet it remains challenging because evasive language is often subtle and highly context-dependent.

SemEval-2026 Task 6, CLARITY (Thomas et al., 2026), formalizes this problem as two classification tasks over the QEvasion dataset (Thomas et al., 2024). **Subtask 1** predicts one of three clarity levels (*Clear Reply*, *Ambivalent*, *Clear Non-Reply*), while **Subtask 2** predicts one of nine fine-grained evasion types. The taxonomy matters because each evasion type belongs to exactly one clarity level, so systems can either exploit or ignore meaningful hierarchical structure. Both subtasks are evaluated with macro F1, which gives equal weight to each class and therefore makes severe imbalance especially consequential (13$\times$ ratio between the largest and smallest classes in Subtask 2).

Our approach deliberately avoids LLM-based prompting strategies and instead asks how far **smaller, fine-tuned encoder models** can go with careful ensembling. We train 50 models per architecture (DeBERTa-xlarge, 900M; DeBERTa-v3-large, 304M) using different random seeds and aggregate them with simple logit averaging. This design philosophy prioritizes **accessibility and reproducibility**: our largest model is under 1B parameters, runs on a single GPU, and requires no API access or proprietary LLM inference, making the system practical for researchers with limited computational budgets. The multi-seed strategy is motivated by the well-documented instability of fine-tuning large pre-trained language models (Dodge et al., 2020), where individual runs can vary by up to 0.10 F1 from random seed differences alone.

Our system achieves macro F1 of **0.76** (Subtask 1) and **0.50** (Subtask 2). Beyond these results, the main scientific contribution of this paper is what we term the **optimization paradox**: three indepen-

*Our official submissions appear under the participant name "taleef" on the Codabench and task page leaderboards.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1518–1525

dent post-hoc calibration techniques, each tuned on out-of-fold (OOF) predictions, consistently improved OOF metrics but degraded evaluation scores by 0.02–0.10 F1. We frame this as suggestive rather than definitive evidence. With only 237 evaluation samples, two of the three degradations lie within the ± 0.06 CI expected from sampling noise, so the observation is interesting primarily because of the *consistent directional pattern* across interventions. We interpret the pattern as a distinction between *model-level* interventions (multi-seed ensembling, which transfers) and *prediction-level* interventions (post-hoc calibration, which appears more prone to overfitting), offering a practical guideline for shared-task settings with limited evaluation data. Our error analysis further shows that Subtask 2 difficulty is driven by semantic overlap between evasion categories and extreme class imbalance, with four minority classes achieving F1 below 0.27.

2 Background

2.1 Task and Dataset

The CLARITY shared task follows the two-level taxonomy from Thomas et al. (2024). At the coarse level, responses are categorized as *Clear Reply*, *Ambivalent*, or *Clear Non-Reply*. At the fine-grained level, nine evasion types (e.g., *Dodging*, *Deflection*, *Implicit*) each map to exactly one clarity level, making the taxonomy both linguistically meaningful and operationally useful for modeling. The QEvasion dataset (Thomas et al., 2024) contains 3,448 English question-answer pairs from political interviews, annotated using a combination of GPT-3.5 and human annotators. The held-out SemEval evaluation set contains 237 samples. Table 1 shows the training distribution: *Ambivalent* dominates Subtask 1 (59.2%), while Subtask 2 is severely long-tailed, with *Partial/half-answer* comprising only 2.3% of samples.

Clarity Level	Evasion Type	N	%
Clear Reply	Explicit	1,052	30.5
Ambivalent	Dodging	706	20.5
	Implicit	488	14.2
	General	386	11.2
	Deflection	381	11.0
	Partial/half-answer	79	2.3
Clear Non-Reply	Declining to answer	145	4.2
	Claims ignorance	119	3.5
	Clarification	92	2.7

Table 1: Training set class distribution (N=3,448) with hierarchical label structure. Five of nine evasion types comprise <5% of samples.

2.2 Related Work

Political evasion has been studied through qualitative frameworks (Bull, 2003; Harris, 1991; Clayman, 2001). Thomas et al. (2024) bridged this with NLP, introducing QEvasion with encoder (DeBERTa/RoBERTa/XLNet) and LLM (ChatGPT, Llama, Falcon) baselines. Our work builds on the DeBERTa family (He et al., 2021, 2023). The multi-seed ensemble approach is motivated by findings that fine-tuning pre-trained Transformers is highly sensitive to random initialization (Dodge et al., 2020; Mosbach et al., 2021) and by the deep-ensembles result that averaging independently trained networks yields robust gains in accuracy and calibration (Lakshminarayanan et al., 2017). Our contribution is empirical: we document the scale of seed-induced variance on a pragmatic classification task (up to 0.10 F1 spread) and characterize the OOF-vs-evaluation behavior of seed ensembling versus post-hoc calibration on a small, imbalanced evaluation set, where such calibration is known to be sensitive to validation composition (Guo et al., 2017).

3 System Overview

3.1 Design Philosophy

We chose fine-tuned encoder models over LLM prompting for three reasons. First, **parameter efficiency**: our largest model (DeBERTa-xlarge, 900M parameters) is 7–70$\times$ smaller than typical prompted LLMs (7B–70B). Second, **reproducibility**: fine-tuning with fixed seeds is fully deterministic, unlike temperature-dependent LLM generation. Third, **task adaptation**: fine-tuning on the full training set lets the model absorb task-specific annotation conventions that are difficult to communicate reliably through prompts alone.

To validate this choice, we evaluated Qwen2.5-32B-Instruct (Yang et al., 2024) in a 5-shot prompting setup (per-class exemplars, greedy decoding, temperature 0) on the training set (N=3,448). Table 2 shows the resulting label distribution. The naively prompted LLM exhibits severe label bias: it massively over-predicts *Dodging* and *Partial/half-answer* while nearly ignoring the most frequent class, *Explicit*. We emphasize that this is a test of *naive 5-shot prompting without chain-of-thought (CoT) or hierarchical decomposition* on a single

Evasion Type	True %	Pred. %
Explicit	30.5	4.5
Dodging	20.5	48.7
Implicit	14.2	1.6
General	11.2	4.9
Partial/half-answer	2.3	34.7
4 others	21.3	5.6

Table 2: Label distribution: Qwen2.5-32B 5-shot predictions vs. true labels. The LLM massively over-predicts *Dodging* and *Partial/half-answer* while ignoring *Explicit*.

model, not a claim about LLMs in general. It supports the narrower conclusion that encoder fine-tuning is a strong, simple baseline relative to naively prompted mid-scale LLMs. Consistent with the official task overview (Thomas et al., 2026), stronger shared-task systems used more capable LLM prompting pipelines and hierarchical exploitation of the taxonomy, and we do not dispute that such systems can surpass encoder ensembles on CLARITY.

3.2 Model Architecture

We use **DeBERTa-xlarge** (900M params, disentangled attention) for Subtask 1 and **DeBERTa-v3-large** (304M params, ELECTRA-style pretraining) for Subtask 2. Both use CLS pooling $\rightarrow$ dropout (p=0.1) $\rightarrow$ linear classifier. Input format: `Question: {q}\nAnswer: {a}`, max 512 tokens. Figure 1 illustrates the system architecture.

3.3 Multi-Seed Ensemble Strategy

We train 50 models per architecture via 5-fold stratified cross-validation $\times$ 10 random seeds. At inference, we average raw logits across all models:

$$\hat{y} = \arg\max_{c} \frac{1}{M} \sum_{m=1}^{M} z_m^{(c)} \qquad (1)$$

where $z_m^{(c)}$ is the logit for class c from model m and M=50. This is motivated by the high variance across seeds: in our DeBERTa-v3-large baseline experiments on Subtask 1 (before selecting xlarge as the final T1 architecture), individual models ranged from 0.601 to 0.698 macro F1 (σ=0.024), which is a spread of nearly 0.10 purely from random initialization.

Subtask 1 and 2 predictions are made **independently**: we do not condition Subtask 2 on Subtask 1 outputs, as hierarchical conditioning amplified cascading errors (Section 5.3).

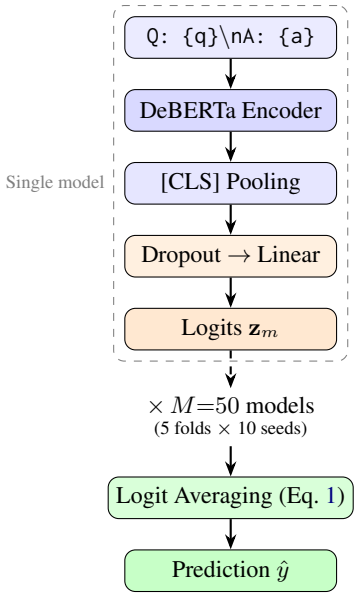

Figure 1: System architecture. Each model independently maps a QA pair through a DeBERTa encoder with a linear classification head. Final predictions aggregate logits from 50 models via simple averaging.

4 Experimental Setup

All models share identical hyperparameters: learning rate 2×10^{-5} with 10% linear warmup, AdamW (Loshchilov and Hutter, 2019) with weight decay 0.01, effective batch size 32 (4 $\times$ 8 gradient accumulation), cross-entropy loss, 3 epochs, using HuggingFace Transformers (Wolf et al., 2020). Training was performed on NVIDIA A100-80GB GPUs (university HPC cluster). Each xlarge model takes $\sim$45 min and v3-large $\sim$15 min, totaling $\sim$50 GPU-hours for all 100 models. Full hyperparameters are listed in Appendix A.1.

5 Results and Analysis

5.1 Official Results and Baselines

Table 3 presents our development baselines, ensemble results, and official evaluation scores. The official task overview reports that our system achieved macro F1 = 0.76 on Subtask 1 and 0.50 on Subtask 2, while the top system reached 0.89 and 0.68 respectively (Thomas et al., 2026). These scores correspond to **18th of 41 participants** on Subtask 1 and **12th of 33 participants** on Subtask 2 on the official CLARITY task page.[1] This gap is consistent with the official task paper's broader finding that the strongest systems generally relied

[1]Rank placement taken from the official CLARITY task page: https://konstantinosftw.github.io/CLARITY-SemEval-2026/

on LLM prompting and hierarchical exploitation of the taxonomy, whereas systems treating the two subtasks independently were generally weaker (Thomas et al., 2026). Our approach intentionally takes the latter path: two independently fine-tuned encoder ensembles, chosen for accessibility, sub-1B scale, single-GPU training, and no API dependence.

For prior work, Table 3 compares against the original QEvasion baselines from Thomas et al. (2024). That paper does not report a plain BERT baseline, so the closest encoder references are DeBERTa-base, RoBERTa-base, and XLNet-base, alongside ChatGPT as the reported LLM baseline.

The 50-model DeBERTa-v3-large ensemble improves over the single-model mean by +0.024 on Subtask 1 OOF (0.667 vs. 0.643) and +0.046 on Subtask 2 OOF (0.373 vs. 0.327). For DeBERTa-xlarge on Subtask 1, our final submission architecture, the 50-model ensemble improves from 0.663 ± 0.021 to 0.679 (+0.016 OOF).

The higher evaluation F1 relative to OOF (T1: 0.76 vs. 0.679, +0.081; T2: 0.50 vs. 0.373, +0.127) reflects a structural difference—OOF models see only $4/5$ of the training data, whereas evaluation predictions aggregate 50 models trained on the full set—compounded on T2 by macro F1's extreme sensitivity under severe imbalance, where a handful of minority-class predictions among 237 samples can swing the score by multiple points. We cannot fully rule out that part of either jump reflects favorable minority-class sampling in the evaluation set, and both numbers should be read with this caveat.

5.2 Ensemble Size Ablation

Figure 2 shows OOF macro F1 as a function of ensemble size, computed by incrementally adding models and re-evaluating on the OOF predictions. For DeBERTa-xlarge on Subtask 1, the xlarge single-model mean is 0.663 ± 0.021 (Table 3), and the 50-model ensemble achieves 0.679; the curve starts at 0.323 for a single fold $\times$ single seed, which reflects extreme per-fold evaluation variance on $\sim$690 samples rather than typical single-model performance. The steepest gains occur in the first 10–15 models, with diminishing returns thereafter, confirming that the ensemble consistently outperforms typical individual runs.

5.3 The Optimization Paradox

A central finding is that *every* post-hoc optimization we attempted improved OOF performance but degraded evaluation performance. Table 4 summarizes this pattern across two categories of intervention, each measured against its appropriate baseline. We view the result as suggestive evidence rather than a universal claim, because the evaluation set is small and therefore noisy.

System	T1	T2
Prior work baselines[†]		
ChatGPT (ZS) (Thomas et al., 2024)	0.413	0.244
DeBERTa-base (FT) (Thomas et al., 2024)	0.441	–
RoBERTa-base (FT) (Thomas et al., 2024)	0.530	–
XLNet-base (FT) (Thomas et al., 2024)	0.518	–
Development baselines (OOF)		
Majority class	0.248	0.052
TF-IDF + LR	0.546	0.319
v3-large (single model)	.643±.024	.327±.040
xlarge (single model)	.663±.021	–
Our ensembles (OOF)		
v3-large (50 models)	0.667	0.373
xlarge (50 models)	0.679	–
Official evaluation		
Single-seed v3-large	0.54	0.50
Multi-seed ensemble	**0.76** (18/41)	**0.50** (12/33)
+ post-hoc optimizations	0.74	0.40

Table 3: Development and official results. [†]Prior work baselines were evaluated on the original QEvasion test split (317 samples) (Thomas et al., 2024), not the SemEval evaluation set (237 samples). Thomas et al. (2024) do not report a plain BERT baseline; the closest encoder baselines are DeBERTa-base, RoBERTa-base, and XLNet-base, with ChatGPT (gpt-3.5-turbo) as the reported LLM baseline. T1 uses direct clarity macro F1; T2 uses 9-class evasion macro F1 (reported only for ChatGPT). "Single model" = mean±std over held-out-fold F1 for all 50 models. Rank in parentheses.

The most striking case is **hierarchical masking**, which learns a 3×9 matrix $\mathbf{M}$ to adjust Subtask 2 logits based on Subtask 1 probabilities: $\mathbf{z}_2' = \mathbf{z}_2 + \mathbf{p}_1\mathbf{M}$. On OOF data, this improved F1 by +0.026, but on evaluation it caused a -0.10 drop, likely because Subtask 1 errors propagated through $\mathbf{M}$ and amplified Subtask 2 errors.

We identify a key distinction: the multi-seed ensemble is a *model-level* intervention (training different models), while all failed optimizations are *prediction-level* interventions (transforming fixed OOF outputs). The former introduces genuine diversity; the latter can only redistribute decisions within a fixed prediction space, making it more prone to overfitting when the OOF distribution differs from evaluation.

With only 237 evaluation samples, the 95% CI for macro F1 is approximately ±0.06: the -0.02 T1 drops individually fall within noise, while the -0.10 T2 drop from hierarchical masking clearly

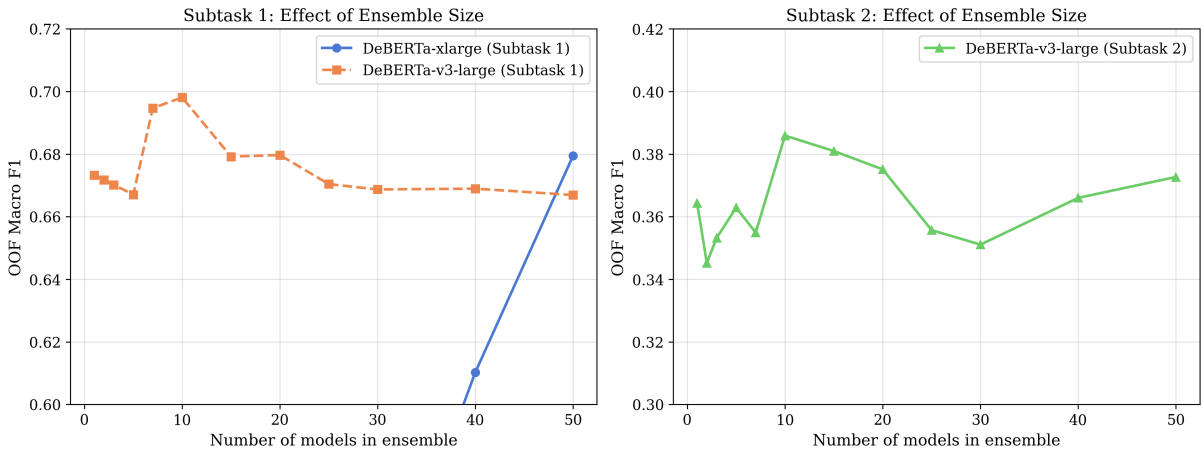

Figure 2: OOF macro F1 vs. ensemble size for both architectures and subtasks. For DeBERTa-v3-large, most gains are achieved within the first 25 models; DeBERTa-xlarge benefits more gradually across the full ensemble, with the largest late-stage gain between 40 and 50 models.

Intervention	**Task**	**OOF Δ**	**Eval Δ**
Model-level (Δ vs. single-model mean)			
Multi-seed ens.	T1	+0.016	**+0.22**
Multi-seed ens.	T2	+0.046	0.00
Prediction-level (Δ vs. multi-seed ensemble)			
Optuna weights (Akiba et al., 2019)	T1	+0.003	−0.02
Class thresholds	T1	+0.007	−0.02
Hier. masking	T2	+0.026	−0.10

Table 4: The optimization paradox. OOF Δ vs. single-model mean (xlarge for T1, v3-large for T2); Eval Δ vs. our single-seed v3-large submission (the only pre-ensemble submission for either subtask). The +0.22 T1 entry therefore conflates architecture (v3-large → xlarge) and ensembling effects and is an upper bound on the pure ensembling contribution; prediction-level rows all use the same ensemble backbone and cleanly isolate calibration effects.

exceeds it. The paradox should therefore be interpreted cautiously. Our main evidence is the consistency of the directional pattern—*every* prediction-level intervention degraded evaluation while model-level ensembling improved it—rather than any claim that each individual drop is statistically decisive. A plausible interpretation is that prediction-level tuning overfits to distributional artifacts in OOF predictions (e.g., class-specific calibration biases) that do not transfer when evaluation class proportions differ, whereas model-level diversity is more robust to such shifts.

5.4 Error Analysis and Task Insights

Subtask 1: The Ambivalent Boundary Problem. As shown in Figure 3 (left), the primary difficulty in Subtask 1 is distinguishing *Ambivalent* from *Clear Reply*: these two classes account for 734 of 986 total OOF errors (74.4%). This confusion reflects a genuine annotation challenge since many political responses begin with a superficially direct statement but embed hedging or topic shifts. For example, the response *"No. I think we're going to be strategic partners"* to a question about U.S.-China relations was annotated as *Ambivalent* (the initial "No" contradicts the subsequent positive framing), but our model predicts *Clear Reply* with 0.605 confidence, likely due to the unambiguous "No" token. This suggests that detecting *pragmatic* evasion (where surface-level directness masks substantive non-answers) remains a core challenge.

Subtask 2: Majority-Class Collapse. Figure 3 and Table 6 (Appendix A.3) reveal that Subtask 2 errors follow a systematic pattern: the model misclassifies minority evasion types as the dominant *Explicit* class (188 *Implicit*, 139 *General*, 115 *Dodging*, and 110 *Deflection* samples). This "majority-class collapse" is most extreme for *Partial/half-answer* (79 training samples, F1=0.00), which is never predicted. The pattern suggests that standard cross-entropy training is insufficient for this extreme imbalance. We implemented focal loss (Lin et al., 2017) (γ=2.0) and inverse-frequency class weighting as alternatives, but did not submit these variants during the evaluation phase: with only three submission slots per subtask and our Subtask 1 ensemble already performing well, we prioritized submission stability over Subtask 2 exploration. Post-acceptance experiments with these loss functions (Appendix A.5) confirm that inverse-

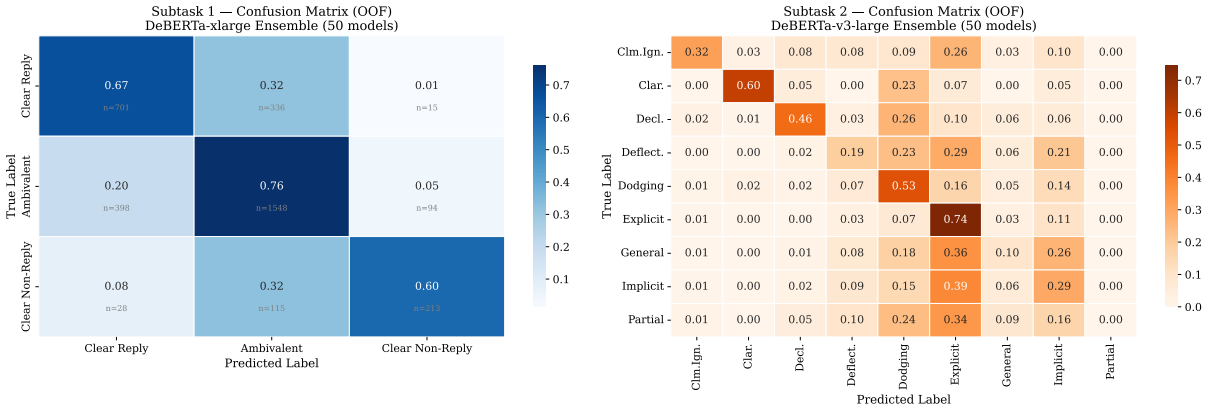

Figure 3: OOF confusion matrices (normalized by true label). Left: Subtask 1 (DeBERTa-xlarge, 50 models), showing strong bidirectional confusion between *Ambivalent* and *Clear Reply*. Right: Subtask 2 (DeBERTa-v3-large, 50 models), showing systematic minority-class collapse into *Explicit*.

frequency class weighting yields a modest +.024 OOF macro F1 gain over plain cross-entropy, while focal loss did not help.

Interestingly, *Clarification* achieves the highest F1 (0.655) despite being a minority class (92 samples), likely because clarification responses have a distinctive linguistic structure (asking questions back). In contrast, *General* (F1=0.136) and *Deflection* (F1=0.234) are semantically close to *Dodging*, differing primarily in the *intent* behind the evasion rather than its surface form, which is a distinction that may require discourse-level reasoning beyond what token-level classifiers can capture.

6 Conclusion

We presented a multi-seed DeBERTa ensemble system for SemEval-2026 Task 6 (CLARITY), achieving macro F1 scores of 0.76 and 0.50 on Subtasks 1 and 2 respectively. We also report suggestive evidence for an *optimization paradox*: model-level ensembling improved Subtask 1 and did not harm Subtask 2, whereas three prediction-level post-hoc calibrations each regressed evaluation performance. Because the evaluation set is limited, this observation should be interpreted cautiously and rests mainly on the consistent directional pattern rather than any individually definitive drop. Even with that caveat, it offers a practical shared-task recommendation: when evaluation data is limited, invest in training more diverse models before optimizing fixed predictions. Our error analysis suggests that the main challenges are the *Ambivalent*/*Clear Reply* boundary in Subtask 1 and majority-class collapse in the long-tailed Subtask 2 distribution.

Limitations

Our multi-seed ensemble requires training 100 models (∼50 GPU-hours), though our ablation shows that 10–15 seeds capture most of the benefit. We did not explore data augmentation or multi-task learning between subtasks, and our LLM comparison is limited to a single model (Qwen2.5-32B). Our optimization paradox analysis rests on three submissions per subtask, and the limited 237-sample evaluation size likely contributes variance.

Ethics Statement

Automated classification of political speech carries both social value and risk. It can support media accountability, debate analysis, and political communication research, but classifier errors, especially the *Ambivalent*/*Clear Reply* confusion we document, can mischaracterize a politician's response and should not be treated as definitive judgments. The model may also inherit biases from training data drawn primarily from U.S. political interviews, so outputs should be validated by human reviewers before use in high-stakes settings.

Acknowledgments

We thank the CLARITY task organizers and Purdue University Fort Wayne's Gilbreth HPC cluster for data and compute support.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, Anchorage, Alaska, USA. ACM.

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

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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, Venice, Italy. IEEE.

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

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations (ICLR)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics. To appear.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A Reproducibility Details

A.1 Training Hyperparameters

Hyperparameter	Value
Learning rate	2×10^{-5}
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	0.10
LR scheduler	Linear
Epochs	3
Batch size	32 (4×8 grad. accum.)
Max sequence length	512 tokens
Dropout	0.1
Loss function	Cross-entropy

Table 5: Training hyperparameters, identical for all 100 models.

A.2 Software Environment

Python 3.10, PyTorch 2.1.0, Transformers 4.36.0 (Wolf et al., 2020), scikit-learn 1.3.2 (Pedregosa et al., 2011), CUDA 12.1, NVIDIA A100-80GB.

A.3 Per-Class Results

Class	P	R	F1
Subtask 1 (DeBERTa-xlarge, 50 models)			
Clear Reply	.622	.666	.643
Ambivalent	.774	.759	.767
Clear Non-Reply	.662	.598	.628
Subtask 2 (DeBERTa-v3-large, 50 models)			
Clarification	.724	.598	.655
Explicit	.554	.744	.635
Dodging	.487	.534	.509
Declining to answer	.545	.462	.500
Claims ignorance	.603	.319	.418
Implicit	.247	.291	.267
Deflection	.294	.194	.234
General	.221	.098	.136
Partial/half-answer	.000	.000	.000

Table 6: Per-class OOF results (held-out folds, ∼3,448 predictions). These OOF values differ from the official evaluation scores (0.76/0.50) because each OOF prediction comes from a model trained on only $4/5$ of the data; see §5.1 for discussion of the OOF-to-eval gap. Subtask 2 minority classes collapse into *Explicit*, with *Partial/half-answer* never predicted.

A.4 Seed Variance

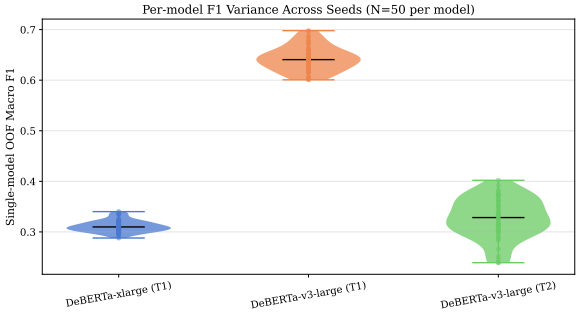

Figure 4: Per-seed OOF macro F1 across all 50 models per architecture, illustrating the fine-tuning instability that motivates multi-seed ensembling.

A.5 Class Imbalance Mitigation

To address Subtask 2 class imbalance, we compare three loss functions using DeBERTa-v3-large with 5-fold CV. Table 7 reports per-fold OOF macro F1.

Loss Function	Macro F1	Min-class F1
Cross-entropy (baseline)	.273 ± .034	.000
Focal loss (γ=2.0)	.260 ± .047	.000
Class-weighted CE	**.297 ± .022**	.000

Table 7: Loss function ablation on Subtask 2 (OOF macro F1, mean±std over 5 folds, single seed, DeBERTa-v3-large). Min-class F1 = F1 of the worst-performing class.

Class-weighted CE provides a modest +0.024 absolute macro F1 improvement over plain CE while reducing fold variance (.022 vs. .034). Inspecting per-class F1, the gain comes from rare-class recovery: Deflection improves from .045 to .193, General from .027 to .129, and Claims ignorance from .230 to .434, at the cost of a $-.21$ drop on the majority Explicit class (.587→.379). Focal loss with γ=2.0 underperforms plain CE in our setting, suggesting that down-weighting easy examples does not transfer benefits from object-detection contexts to this 9-class evasion task. Critically, all three losses produce zero F1 on *Partial/half-answer* (21 evaluation samples per fold on average), indicating that ultra-rare classes require either data augmentation or hierarchical decomposition rather than re-weighting alone. These results suggest class weighting is the most promising single-fix mitigation, but full closure of the imbalance gap likely requires methods beyond loss reshaping.