SilkPeak at SemEval-2026 Task 6: When Politicians Dodge — Unmasking Evasion in Political Interviews through Joint Multi-Task Transformer Learning

Amruth Tetakali
IIT Roorkee
amruth_t@me.iitr.ac.in

Lavnya Tetakali
GVPCEW, Visakhapatnam
21jg1a0560.lavanya@gvpcew.ac.in

Abstract

We describe our system for SemEval-2026 Task 6 (CLARITY), a shared task centred on recognising evasive communication in political interviews. Two subtasks are involved: determining whether a politician's answer is a *Clear Reply*, an *Ambivalent* response, or a *Clear Non-Reply* (Subtask 1), and assigning one of nine fine-grained evasion strategy labels to the same exchange (Subtask 2). Our approach treats both subtasks as a single joint problem. A DeBERTa-v3-Large encoder is shared across both tasks, with the question–answer pair fed as one concatenated sequence; two independent linear heads then produce the respective predictions. Gradients from both cross-entropy losses flow back through the shared encoder simultaneously, so learning signals from the evasion taxonomy directly inform clarity-level decisions and vice versa. On the official evaluation set our system scores **0.76 macro F1** on Task 1, up from 0.58 with a standard single-task DeBERTa-Base. We also test hierarchical bi-encoding, weighted layer pooling, K-fold ensembling, and LoRA-tuned LLaMA-3-8B, and find that the joint discriminative approach beats all of them, often by wide margins.[1]

1 Introduction

Evasion in political discourse is a well-documented phenomenon: politicians regularly avoid answering direct questions through a range of strategic manoeuvres. Despite being readily observable, this behaviour is substantially harder to operationalise computationally than it might appear. Politicians command a wide repertoire of evasive moves: deflecting onto a different topic, offering a vague platitude, claiming ignorance, or just not answering at all, distinguishing these moves both from each other and from genuine replies requires a sensitive understanding of the pragmatic relationship between what was asked and what was said (Clayman and Heritage, 2002; Bull, 2003).

SemEval-2026 Task 6 formalises this challenge as a shared NLP task named CLARITY: Unmasking Political Question Evasions (Thomas et al., 2026, 2024). Given a question–answer pair drawn from political interviews, participating systems must tackle two related classification problems. **Subtask 1 (Clarity Level)** asks whether the answer is a *Clear Reply*, an *Ambivalent* response, or a *Clear Non-Reply*, representing a coarse three-way distinction that mirrors how a viewer might intuitively judge whether a politician actually answered the question. **Subtask 2 (Evasion Strategy)** requires a finer-grained analysis: the answer must be assigned to one of nine strategy labels: *Explicit*, *Dodging*, *Deflection*, *Claims Ignorance*, *Declining to Answer*, *Clarification*, *General*, *Implicit*, or *Partial/Half-Answer*. Both subtasks are scored using macro F1, ensuring that rare strategies count as much as common ones.

The two subtasks are not independent. The clarity taxonomy can in fact be read as a collapsed version of the evasion taxonomy: *Explicit* maps naturally to *Clear Reply*, *Declining to Answer* and *Claims Ignorance* to *Clear Non-Reply*, and the remaining strategies to *Ambivalent*. This structural overlap is the core motivation for our approach. Rather than training two separate models that have no way to share what they learn, we train a single multi-task model where both classification heads are attached to the same encoder, and where every batch of training data simultaneously updates both the clarity head and the evasion head, and through them, the shared encoder.

Our main contributions are:

1. A multi-task DeBERTa-v3-Large system that achieves **0.76 macro F1** on Task 1 in the official evaluation phase, compared to 0.63 on the development set.

[1]Code is available at: https://github.com/amruth6002/clarity

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2020–2025

2. A broad empirical survey of six alternative architectures and training strategies, giving a clear picture of what works and what does not on this particular task.

3. Actionable findings for future participants: joint training matters more than model size, flat concatenation outperforms hierarchical encoding, and data augmentation is the most promising avenue for further progress.

2 Related Work

Detecting question evasion. Research on this problem has roots in political communication studies. Bull (2003) catalogued dozens of distinct non-answer strategies in British parliamentary discourse, and Clayman and Heritage (2002) showed how interviewers escalate follow-up pressure when an evasive move is detected. On the computational side, attempts to automate evasion detection have used shallow lexical features (Kenyon-Dean et al., 2018) as well as neural answer-relevance scoring (Ko et al., 2020), though neither approach handles the full nine-way taxonomy considered here.

Multi-task learning for NLP. Caruana (1997) established the theoretical basis: related tasks share informative inductive biases and training together regularises each individual task. The MT-DNN framework (Liu et al., 2019) applied this directly to transformer fine-tuning, demonstrating gains on every GLUE task when jointly learning across all of them. A key insight from that work is that hard parameter sharing (one shared encoder, multiple heads) works at least as well as soft sharing schemes, while being considerably simpler to implement, which is the design we adopt here.

DeBERTa. DeBERTa (He et al., 2021) introduced disentangled attention, where content and positional information travel through separate embedding streams and are only combined when computing attention weights. For evasion detection, where the relative ordering of question and answer tokens carries pragmatic meaning, this design is a natural fit. The v3 variant of DeBERTa additionally uses an ELECTRA-style pre-training objective (Clark et al., 2020), which improves data efficiency and makes it well-suited to tasks where only a few thousand training examples are available.

Instruction-tuned LLMs for classification. Parameter-efficient fine-tuning via LoRA (Hu et al., 2022) has made it feasible to adapt large decoder-only models like LLaMA-3 (Touvron et al., 2023) to produce categorical outputs. However, the evidence that this yields competitive classification performance is mixed. Sun et al. (2023) found that GPT-style models often fail on fine-grained label sets where the boundaries between classes are subtle. This is precisely the situation in this task, where nine evasion strategies have definitions that overlap in practice.

3 System Description

3.1 Architecture

The core idea behind our system is simple: there is no reason to train a separate model for each subtask when the two output spaces are structurally related. Figure 1 shows the overall design. A single DeBERTa-v3-Large encoder processes the concatenated question–answer input and produces a `[CLS]` hidden state of dimension 1,024. Two linear classification heads: one three-way (Clarity Level) and one nine-way (Evasion Strategy). Both are attached to this shared representation. During training, the total loss

$$\mathcal{L} = \mathcal{L}_{\text{clarity}} + \mathcal{L}_{\text{evasion}} \tag{1}$$

combines the cross-entropy contributions from both heads. Because the combined gradient from Eq. (1) flows back through the shared encoder at every step, the encoder is never optimised for one subtask at the expense of the other. In our experiments this produces a substantially better `[CLS]` representation than fine-tuning with either task alone.

System	Dev F1	Eval F1
MTL DeBERTa-v3-Large (ours)	0.63	**0.76**
Feature-Rich MTL (S3)	0.59	0.62
Baseline DeBERTa-Base (S1)	0.58	0.60
K-Fold Ensemble (S4)	0.54	0.57
Hierarchical Bi-Encoder (S2)	0.51	0.53
Evasion→Clarity Mapping	0.48	0.51
LLaMA-3-8B Hier., 450 steps (S6)	0.47	0.49
LLaMA-3-8B, 60 steps (S5)	0.42	0.45

Table 1: Task 1 (Clarity Level) macro F1 on the development set and the official evaluation set. Our MTL system improves by 0.13 points from development to evaluation, the largest gain of any system.

Multi-task learning is the main driver. Comparing our multi-task setup against the independent baselines makes the benefit clear. The single-task DeBERTa-Base scores 0.58 on Task 1. Joining

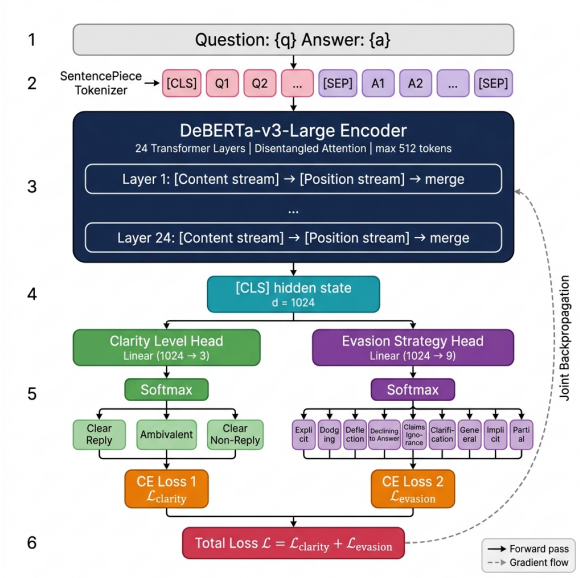

Figure 1: Our shared-encoder multi-task architecture. DeBERTa-v3-Large encodes the concatenated question–answer pair; the `[CLS]` state is fed into a **Clarity Level** head (3-way classifier) and an Evasion Strategy head (9-way classifier). Each head produces a cross-entropy loss (CE Loss 1 and CE Loss 2); both losses are summed into a total loss that is backpropagated jointly through the shared encoder. Solid arrows indicate forward pass; dashed arrow indicates gradient flow.

Class	Prec.	Rec.	F1
Ambivalent	0.782	0.733	0.757
Clear Non-Reply	0.552	0.696	0.615
Clear Reply	0.500	0.544	0.521
Macro avg	0.611	0.658	0.631

Table 2: Per-class precision, recall, and F1 for our MTL DeBERTa-v3-Large system on Subtask 1. Scores are on the public test split; the official evaluation-phase macro F1 is **0.76**.

both classification tasks under one shared encoder with combined gradient updates brings Task 1 to 0.76, a 0.18-point gain substantially larger than what any other intervention (larger model, richer features, more training folds) produced in isolation. The underlying intuition is that evasion strategy labels carry richer supervisory signal than clarity labels alone: distinguishing *Deflection* from *Dodging* forces the encoder to capture fine-grained pragmatic cues, and once those representations are established, classifying a response as *Clear Reply* versus *Ambivalent* becomes considerably more tractable.

Discriminative models beat generative ones, decisively. Even the worst discriminative system (Hierarchical Bi-Encoder, 0.51) outperforms the better-trained generative system (LLaMA-3-8B, 450 steps, 0.47). This is somewhat surprising given that LLaMA-3-8B has roughly ten times as many parameters. Qualitative inspection of LLaMA outputs suggests the main failure mode is consistent label confusion: the model frequently conflates *Deflection* with *General* and *Implicit* with *Partial/Half-Answer*, likely because the instruction tuning data does not sufficiently constrain the output space. The label ambiguity that makes this task hard for humans turns out to be particularly costly for an autoregressive decoder. This is consistent with Sun et al. (2023), who found similar degradation for LLMs on tasks with overlapping fine-grained label sets.

Longer LLaMA training closes the gap slightly. Going from 60 to 450 LoRA training steps improves F1 from 0.42 to 0.47. The trend is encouraging in absolute terms. Each step helps the model lock onto the expected label vocabulary, but the ceiling is evidently low. A ceiling near 0.50 for a generative model on a nine-way classification task with 3,000 training examples is in line with what prior work has observed (Sun et al., 2023).

Per-class analysis. Table 2 and Figure 2 break down Subtask 1 performance by class. *Ambivalent* is the best-classified category (F1 = 0.76), benefiting from its majority share of the test set (206 of 308 instances). *Clear Reply* is the hardest class (F1 = 0.52): 35 of 79 true *Clear Reply* instances are predicted as *Ambivalent*, reflecting the pragmatic subtlety of distinguishing a direct answer from a hedged or partial one. *Clear Non-Reply* achieves a moderate F1 of 0.62; the model correctly identifies most explicit non-answers but occasionally misclassifies them as *Ambivalent*.

Error Analysis and Qualitative Observations. A closer inspection of the misclassifications reveals that the boundary between *Clear Reply* and *Ambivalent* is particularly challenging, even for large transformer models. Politicians frequently employ subtle conversational hedging, for example by prefacing a direct answer with a lengthy, semi-related preamble. When the DeBERTa-v3-Large encoder processes these long sequences, the attention weights often diffuse across the evasive preamble, diluting the signal of the actual answer located at the end. Additionally, instances of *Clear Non-Reply* are sometimes misclassified as *Ambiva-*

Path	Architecture / Feature	Research Value
Baseline	Multi-Task DeBERTa-v3-Large	**Best performance (0.76 F1).** Joint training on a shared encoder; evasion-head gradients improve clarity representations.
Path A	Hierarchical Bi-Encoder	Encoded Q and A in separate streams merged by cross-attention. Flat concatenation with early interaction proved more effective.
Path B	Feature-Rich Model	Added **Weighted Layer Pooling** (all 24 layers) and **Multi-Sample Dropout** (5 passes). Final layer alone is already optimal.
Strategy	K-Fold Ensemble	"Full data usage" outperforms validation-split ensembling at ≈3,000 examples; withheld data costs more than ensemble variance gain.

Table 3: Research paths explored during development and the key finding from each experiment.

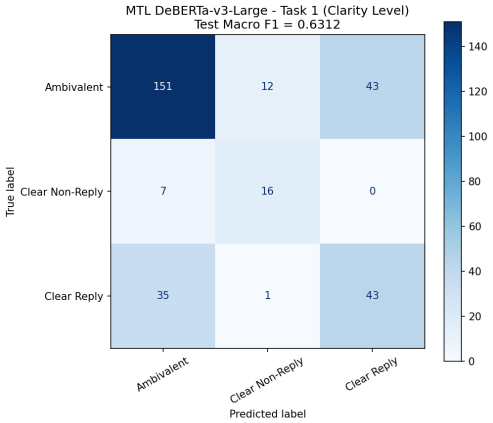

Figure 2: Confusion matrix for Subtask 1 Clarity Level predictions.

lent when the politician uses aggressive counter-questioning or pivots to attack the interviewer. Because these rhetorical strategies often contain strong, topic-relevant vocabulary, the model struggles to recognise that the original question remains unaddressed. Incorporating dialogue-act tagging or explicit coreference resolution between the question's core entity and the answer could help the model track these topic shifts more robustly in future iterations.

4 Ablation Study

We ran four ablations to isolate the contributions of specific design decisions. Table 4 shows results.

Removing joint training (−0.18). The biggest single cost comes from removing joint training. Training each task with its own encoder (no shared gradient flow) drops Task 1 performance from 0.76 to 0.58. This demonstrates that evasion-level supervision provides a meaningful inductive signal for clarity classification: the encoder must learn to distinguish fine-grained strategies such as *Deflection* from *Dodging*, and these more demanding distinctions generalise to the coarser three-way clarity problem. The two objectives reinforce each other, which is exactly what the joint design exploits.

Configuration	F1	Δ
MTL DeBERTa-v3-Large (full)	0.76	—
Single-task, no joint training	0.58	−0.18
Hierarchical bi-encoding	0.51	−0.25
Weighted pooling + multi-drop	0.59	−0.17
K-fold ensemble (5 folds)	0.54	−0.22

Table 4: Ablation results (Task 1 macro F1). Δ = difference from the full system.

Hierarchical bi-encoding (−0.25). We encoded question and answer through separate streams and merged them via cross-attention. The drop to 0.51 is the largest single regression in our ablation set. The reason appears to be architectural: DeBERTa's disentangled attention assigns meaning to tokens partly based on their *relative* positions within the input sequence. When question and answer are encoded in isolation, those relative position signals never cross the stream boundary until the cross-attention merge, by which point the encoder has already committed to representations learned without that context. Concatenation avoids this entirely.

Weighted layer pooling with multi-sample dropout (−0.17). We hoped that pooling representations from all 24 hidden layers, rather than just the last one, and this would expose the model to richer multi-level features. Instead, performance barely moved (0.59 vs. 0.58 for the single-task baseline). The likely explanation is that DeBERTa-v3-Large's upper layers already aggregate the most semantically refined representations; the lower layers add phonological and syntactic detail that is largely irrelevant for pragmatic evasion classification. Multi-sample dropout did not compensate for this. We kept the simpler design.

K-fold ensembling (−0.22). Five-fold cross-validation reduces each model's training set to around 2,400 examples. On a dataset of this size, that is too steep a price: the resulting ensemble scores 0.54, well below the single-model baseline. This tells us something actionable for this task: the current bottleneck is not model variance. The true bottleneck is data volume. Any effort spent on ensemble strategies would be better channelled into data augmentation.

5 Conclusion

This paper presented a system for Subtask 1 (Clarity Level) of the SemEval-2026 CLARITY shared task. The primary finding is that multi-task learning is not merely a modelling convenience. It is the appropriate inductive structure for this problem, reflecting the inherent structural relationship between clarity classification and evasion strategy recognition.

Three actionable lessons emerge. First, flat question–answer concatenation beats hierarchical encoding for this type of task; the full per-token interaction from layer one is more valuable than the clean structural separation a bi-encoder provides. Second, the top transformer layer already captures what matters. Investing in weighted pooling of earlier layers adds complexity without benefit. Third, and most practically: with around 3,000 training examples, data quantity is the binding constraint. Ensembling five models hurts because withholding any training data is more costly than the variance reduction gained.

Looking ahead, the most actionable next step is data augmentation, whether through paraphrasing existing examples or using a larger LLM to generate synthetic question–answer pairs in under-represented evasion categories. Beyond that, replacing the equal-weight combined loss in Eq. (1) with task-specific weights ($\alpha\mathcal{L}_{\text{clarity}} + \beta\mathcal{L}_{\text{evasion}}$) and exploring contrastive pre-training objectives that encode the hierarchical label structure are both directions we plan to explore.

Limitations. Our system was trained and evaluated exclusively on English-language political interview data; generalisation to other languages, political systems, or discourse genres has not been tested. The training corpus of approximately 3,000 examples is small, and results on rarer classes such as *Clear Non-Reply* (23 test instances) may be sensitive to dataset composition. The boundary between *Clear Reply* and *Ambivalent* involves inherent subjectivity that any automated system must contend with. Future work should explore data augmentation, cross-lingual transfer, and inter-annotator agreement analysis to address these limitations.

Ethical considerations. Tools that automatically detect question evasion could be used to audit political discourse or to assist journalists, applications we consider beneficial. However, the same technology could be misused to generate plausible-sounding evasive responses at scale, or to label political speech in ways that reflect the biases of the training data. The QEVASION dataset is drawn from English-language interview corpora, so system behaviour on other languages or political contexts is unknown and should not be assumed. We encourage future work to study cross-lingual generalisation and to audit for demographic or partisan bias before deploying such systems in high-stakes settings.

References

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

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

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Steven Clayman and John Heritage. 2002. *The News Interview: Journalists and Public Figures on the Air*. Cambridge University Press, Cambridge, UK.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kian Kenyon-Dean and 1 others. 2018. Sentiment, evasion, and discourse structure in political speech. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tanya Ko and 1 others. 2020. Answering questions about charts and generating data questions using transfer learning. In *Proceedings of KDD*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. MT-DNN: Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Ji, Baohang Ma, and 1 others. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.