KCLarity at SemEval-2026 Task 6: Encoder and Zero-Shot Approaches to Political Evasion Detection

Archie Sage*
King's College London
archie.sage@kcl.ac.uk

Salvatore Greco*
King's College London
salvatore.greco@kcl.ac.uk

Abstract

This paper describes the KCLarity team's participation in CLARITY, a shared task at SemEval 2026 on classifying ambiguity and evasion techniques in political discourse. We investigate two modelling formulations: (i) directly predicting the clarity label, and (ii) predicting the evasion label and deriving clarity through the task taxonomy hierarchy. We further explore several auxiliary training variants and evaluate decoder-only models in a zero-shot setting under the evasion-first formulation. Overall, the two formulations yield comparable performance. Among encoder-based models, RoBERTa-large achieves the strongest results on the public test set, while zero-shot GPT-5.2 generalises better on the hidden evaluation set.

1 Introduction

Public scrutiny of politicians depends not only on access to questioning but also on the clarity of their responses. Prior work shows that politicians exhibit significantly lower clear reply rates than non-politicians during televised interviews (Bull, 2003). Such unclear responses are commonly described as *equivocation* or *evasion* in the political communication literature (Watzlawick et al., 2011; Bavelas et al., 1988). These findings motivate the development of effective and low-cost automated methods for identifying unclear or evasive answers.

The CLARITY (Thomas et al., 2026) shared task at SemEval 2026 focuses on developing natural language processing (NLP) methods for detecting and classifying response ambiguity and evasion strategies in political discourse. The shared task consists of two subtasks: predicting response clarity (Task 1) and identifying evasion strategies (Task 2).

In this paper, we present the systems developed by the KCLarity team for the CLARITY shared task. We evaluate fine-tuned encoder-based models and decoder-only models prompted in a zero-shot setting. For the encoder models, we consider two training targets: (i) predicting clarity labels directly (*direct clarity*) and (ii) predicting evasion labels and inferring clarity via the hierarchical taxonomy (*evasion-based clarity*). We further explore training configurations including per-class loss weighting (Section 3.2), data-splitting strategies (Section 4.1), input representations (Section 3.3), person-name masking (Appendix B), and additional exploratory experiments (Appendix G). For the decoder models, we evaluate zero-shot prediction with both open-weight and commercial models.

Table 1: Taxonomy of response clarity classification.

Clarity Level	Evasion Level
Clear Reply	Explicit
Ambivalent Reply	Implicit, Dodging, General, Deflection, Partial
Clear Non-Reply	Declining, Ignorance, Clarification

On the public test split, our strongest encoder-based model - RoBERTa-large trained to predict evasion labels and mapped to clarity via the taxonomy - performs competitively on both tasks and outperforms the decoder-only models in our zero-shot evaluation. Among zero-shot systems, GPT-5.2 is the strongest, and we use the same evasion-first formulation for consistency. In the official shared task evaluation on the hidden test set, our top submission is the zero-shot GPT-5.2 system, ranking 22nd out of 44 in Task 1 (macro F1 = 0.74) and 13th out of 33 in Task 2 (macro F1 = 0.50).[1]

2 Background

2.1 Task Definition and Dataset

The CLARITY task comprises two subtasks in English political discourse: predicting the response's

*Equal contribution.

[1]Our implementation code is available at: https://github.com/semeval-2026-kclarity/clarity

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3258–3273

clarity level (Task 1) and identifying the evasion technique (Task 2). The tasks are hierarchically related through the taxonomy introduced in the QEvasion dataset (Thomas et al., 2024). The mapping between the two label sets is shown in Table 1.

The QEvasion dataset comprises 3,448 training instances and 308 test instances, drawn from US presidential interviews.[2] Each instance is annotated with a single clarity label. For the evasion level, training instances receive one evasion label, while test instances contain three evasion labels assigned independently by separate annotators. This multi-annotator supervision in the test set informs the evaluation strategy described in Section 4.2. In addition to these primary labels, each instance includes further metadata, such as the interview date, the president, whether the question contains multiple sub-questions, and whether the question is affirmative.

2.2 Class Imbalance and Inter-annotator Agreement

As noted by the shared task organisers, clarity-level inter-annotator agreement (IAA), measured using Fleiss' Kappa κ (Fleiss, 1971) on the QEvasion dataset shows that annotators (i) almost never confuse *Clear Reply* with *Clear Non-Reply* ($\kappa = 0.97$), (ii) sometimes disagree between *Clear Reply* and *Ambivalent* ($\kappa = 0.65$), and (iii) also sometimes disagree between *Clear Non-Reply* and *Ambivalent* ($\kappa = 0.71$) (Thomas et al., 2024). This indicates that the main modelling challenge at clarity-level classification lies in the overrepresented *Ambivalent* class, which accounts for 59.2% of samples, compared with *Clear Reply* and *Clear Non-Reply* at 30.5% and 10.3% prevalence. Accordingly, we explore loss weighting to mitigate the resulting class imbalance (Section 3.2) amongst other strategies.

3 System Overview

We evaluate two approaches for the CLARITY task: (i) fine-tuning encoder-based models and (ii) prompting decoder-only models in a zero-shot setting (Section 3.1). For the encoder models, we explore loss-weighting strategies (Section 3.2), alternative input representations (Section 3.3), and two prediction targets: predicting clarity and evasion separately, or predicting evasion and inferring clarity via the label hierarchy (Section 3.4).

[2]An example question–answer pair is shown in Appendix A.

3.1 Models

Encoder-based models. We fine-tuned RoBERTa (Zhuang et al., 2021) and DeBERTa-v3 (He et al., 2023), evaluating base and large variants. We also conducted preliminary single-seed experiments with BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), and other encoder models, but report results only for RoBERTa and DeBERTa-v3, as the remaining models were consistently less competitive.

Decoder-based models. We prompted in zero-shot settings (Dong et al., 2024) decoder-only models from different families and sizes, including open-weight models such as Llama 3 (Grattafiori et al., 2024; AI@Meta, 2024), Qwen (Team, 2025b), Gemma 3 (Team, 2025a), and a commercial model GPT-5.2[3] (Singh et al., 2025).

3.2 Loss Weighting

To mitigate class imbalance in QEvasion (Section 2.2), we train encoder models with weighted cross-entropy (CE), defined as follows:

$$\mathcal{L}_{\text{WCE}}(x, y) = -\, w_y \log p(y \mid x)$$

Additionally, we compare three weighting schemes. (i) *Unweighted*: $w_y = 1$ for all classes. (ii) *Balanced*: the standard inverse-frequency weight $w_y = \frac{N}{C \cdot n_y}$, where n_y is the class count, N the total number of training instances, and C the number of classes. (iii) *Sqrt*: a milder reweighting $w_y = \frac{1}{\sqrt{f_y + \epsilon}}$, where $f_y = n_y/N$ is the class frequency and ϵ a small stability constant. For the sqrt scheme, weights are capped and rescaled to unit mean to prevent extreme upweighting of rare classes. All weights are computed from and applied only to the training split.

3.3 Input Representation

Each instance consists of a question-answer pair. We compare two encoder input formats that differ in both field ordering and boundary representation.

Segmented representation (answer first). The answer and question are encoded as two segments, with the answer provided first:

`[CLS] a [SEP] q [SEP].`

When supported, token-type embeddings indicate the segment boundary.

[3]Snapshot: `gpt-5.2-2025-12-11`

Table 2: **RoBERTa models: direct vs evasion-based clarity on the public test split.** Macro-F1 (F1), precision (P), and recall (R), averaged over three seeds (standard deviation in brackets). Higher scores between the two formulations are shown in bold.

Model	*Direct clarity*			*Evasion-based clarity*		
	F1	**P**	**R**	**F1**	**P**	**R**
base	**0.598** (0.010)	0.586 (0.013)	**0.636** (0.021)	0.595 (0.028)	**0.589** (0.035)	0.619 (0.030)
large	0.658 (0.024)	0.681 (0.081)	**0.661** (0.030)	**0.661** (0.022)	**0.694** (0.013)	0.641 (0.023)

Marked representation (question first). Alternatively, the question and answer are concatenated into a single sequence with explicit marker tokens:

```
[QUESTION] q [ANSWER] a.
```

Here, the boundary between the two texts is indicated by learned special tokens. We add `[QUESTION]` and `[ANSWER]` as special tokens and resize the model's input embeddings accordingly.

3.4 Direct vs Evasion-Based Clarity

In line with the terminology in the dataset paper (Thomas et al., 2024), we evaluate clarity either *directly* (predicting the three clarity labels) or *via evasion* (predicting the nine-way evasion labels and then mapping predictions to clarity using the taxonomy). Motivated by prior findings that evasion-based clarity can outperform direct clarity (Thomas et al., 2024), we include an ablation on comparable RoBERTa-base and RoBERTa-large models.

Table 2 shows mixed evidence on stability: direct clarity exhibits lower variance for the base model, while for the large model evasion-based clarity is comparably or more stable, particularly in precision. Performance differences are small: direct clarity tends to yield slightly higher recall, while evasion-based clarity yields slightly higher precision; macro F1 is within variance for both settings. Given these limited differences, we focus the remainder of our analysis on evasion-based clarity because (i) it performs comparably to direct clarity, and (ii) it enables a single model trained for evasion to be reused to obtain clarity labels via the mapping, eliminating the need to train a separate model for clarity prediction.

4 Experimental Setup

4.1 Data Splits

We evaluate models under two splitting regimes. In the *label-stratified* setting, we preserve class distributions across training and validation; for Task 2, we apply *dual stratification* over both evasion and mapped clarity labels. In the *president-disjoint* setting, all responses from a given president appear exclusively in one split, preventing speaker leakage and testing cross-speaker generalisation.

Unless otherwise stated, all encoder-based models use an 80/20 dual-stratified split of the 3,448 training samples (2,758 training / 690 validation) and are evaluated on the 308 publicly available test instances. Decoder-only models are evaluated zero-shot on the public test set, so no train–validation split is applied to them. Official rankings are based on the shared task's hidden test set, with predictions submitted via CodaBench[4] to ensure fully blind evaluation. An ablation comparing the two splitting strategies is reported in Appendix E.

4.2 Evaluation Metrics

For clarity level classification (Task 1), each test instance carries a single gold label; we report macro-averaged F1 (F1), Precision (P), and Recall (R).

For evasion level classification (Task 2), each test instance is independently labelled by three annotators. Rather than collapsing this multi-annotator supervision via majority vote - an approach that can discard valuable signal from legitimate disagreement (Fleisig et al., 2023; Basile et al., 2021) - we retain all three annotations. We compute macro-F1 separately against each annotator ($F1_{A1}$, $F1_{A2}$, and $F1_{A3}$) and their average ($F1_{avg}$). We additionally report ACC_{match}, the fraction of predictions matching at least one annotator's label, capturing the assumption that each annotation constitutes a plausible interpretation.

5 Results

In this section, we first present development-phase results for the fine-tuned encoder models (Section 5.1) and zero-shot decoder models (Section 5.2). We then discuss the official system rankings (Section 5.3) and error analysis (Section 5.4).

5.1 Fine-Tuned Encoder Results

Table 3 reports results for RoBERTa and DeBERTa-v3 (base and large), which were fine-tuned to pre-

[4] https://www.codabench.org/

Table 3: **Fine-tuned encoder results on the public test set.** For evasion-based clarity we report macro-F1 (F1), precision (P), and recall (R). For evasion we report $\mathrm{ACC}_{\mathrm{match}}$, the fraction of predictions matching at least one annotator, per-annotator macro-F1, and the average macro-F1 across annotators. Results are averaged over three seeds (mean on the main row, standard deviation in brackets on the row beneath). Best-performing metrics are shown in bold.

Model	*Evasion-based clarity*			*Evasion*				
	F1	**P**	**R**	$\mathbf{ACC_{match}}$	$\mathbf{F1_{A1}}$	$\mathbf{F1_{A2}}$	$\mathbf{F1_{A3}}$	$\mathbf{F1_{avg}}$
RoBERTa-base	0.595	0.589	0.619	0.516	0.349	0.332	0.333	0.338
	(0.028)	(0.035)	(0.030)	(0.014)	(0.024)	(0.007)	(0.008)	(0.011)
RoBERTa-large	**0.661**	**0.694**	**0.641**	0.539	**0.363**	**0.378**	**0.374**	**0.371**
	(0.022)	(0.013)	(0.023)	(0.011)	(0.030)	(0.035)	(0.047)	(0.037)
DeBERTa-v3-base	0.535	0.564	0.565	0.505	0.304	0.337	0.303	0.314
	(0.067)	(0.033)	(0.030)	(0.027)	(0.014)	(0.016)	(0.015)	(0.011)
DeBERTa-v3-large	0.616	0.634	0.610	**0.541**	0.321	0.305	0.325	0.317
	(0.018)	(0.021)	(0.031)	(0.007)	(0.006)	(0.010)	(0.006)	(0.006)

Table 4: **Zero-shot results on the public test set.** For evasion-based clarity we report macro-F1 (F1), precision (P), and recall (R). For evasion we report $\mathrm{ACC}_{\mathrm{match}}$, the fraction of predictions matching at least one annotator, per-annotator macro-F1, and the average macro-F1 across annotators. Best-performing metrics are shown in bold.

Model	*Evasion-based clarity*			*Evasion*				
	F1	**P**	**R**	$\mathbf{ACC_{match}}$	$\mathbf{F1_{A1}}$	$\mathbf{F1_{A2}}$	$\mathbf{F1_{A3}}$	$\mathbf{F1_{avg}}$
gemma_3_27b_it	0.413	0.501	0.446	0.344	0.137	0.149	0.137	0.141
gpt_oss_120b	0.354	0.374	0.354	0.357	0.104	0.121	0.115	0.114
Llama-3.1-8B-Instruct	0.346	0.385	0.370	0.351	0.054	0.079	0.068	0.067
Llama-3.3-70B-Instruct	0.532	0.544	0.547	0.416	0.284	0.292	0.294	0.290
Qwen3-8B	0.337	0.361	0.350	0.292	0.060	0.078	0.070	0.069
Qwen3-32B	0.338	0.383	0.347	0.331	0.063	0.098	0.089	0.083
GPT-5.2	**0.626**	**0.600**	**0.670**	**0.481**	**0.363**	**0.339**	**0.371**	**0.358**

dict evasion labels with clarity inferred via the taxonomy. All models achieve moderate performance, with clarity macro F1 ranging from 0.535 to 0.661 and average evasion F1 from 0.314 to 0.371. RoBERTa-large is the best-performing model on both tasks (clarity F1 = 0.661; evasion $\mathrm{F1}_{\mathrm{avg}}$ = 0.371), approaching the fine-tuned LLaMA-70B baseline (F1 = 0.68) reported by Thomas et al. (2024). DeBERTa-v3-large is the second-best model and more stable across seeds; it achieves the highest $\mathrm{ACC}_{\mathrm{match}}$ (0.541), though this falls within RoBERTa-large's variance. Encoder models generally outperformed most zero-shot decoder models during development (Section 5.2).

Several auxiliary strategies were explored but yielded no substantial or consistent gains: person-name masking (Appendix B), loss weighting via inverse-frequency and sqrt-based schemes (Appendix C), intermediate fine-tuning on Earnings Calls Q&A (Nuaimi et al., 2025) for cross-domain transfer (Appendix G.1), and augmenting inputs with cognitive distortion (CD) probability buckets (Appendix G.2). The latter two approaches additionally introduced training instability.

5.2 Zero-Shot Results

Table 4 reports zero-shot results for the decoder models, all prompted to predict the evasion level with clarity inferred via the taxonomy. The same prompt was used for all models (Appendix K). Among open-weight models, Llama-3.3-70B-Instruct performs best (clarity F1 = 0.532; evasion $\mathrm{F1}_{\mathrm{avg}}$ = 0.290), while smaller models across all families cluster well below. The commercial model GPT-5.2 substantially outperforms all open-weight alternatives on both tasks, achieving a clarity F1 of 0.626 and an evasion $\mathrm{F1}_{\mathrm{avg}}$ of 0.358 with an $\mathrm{ACC}_{\mathrm{match}}$ of 0.481. Nevertheless, all zero-shot decoder models underperform the best encoder models (Section 5.1), with GPT-5.2 being the only model approaching comparable performance.

5.3 Official Ranking

Official rankings were determined on the hidden evaluation set via CodaBench, as described in Section 4.1. We submitted two systems for both tasks:

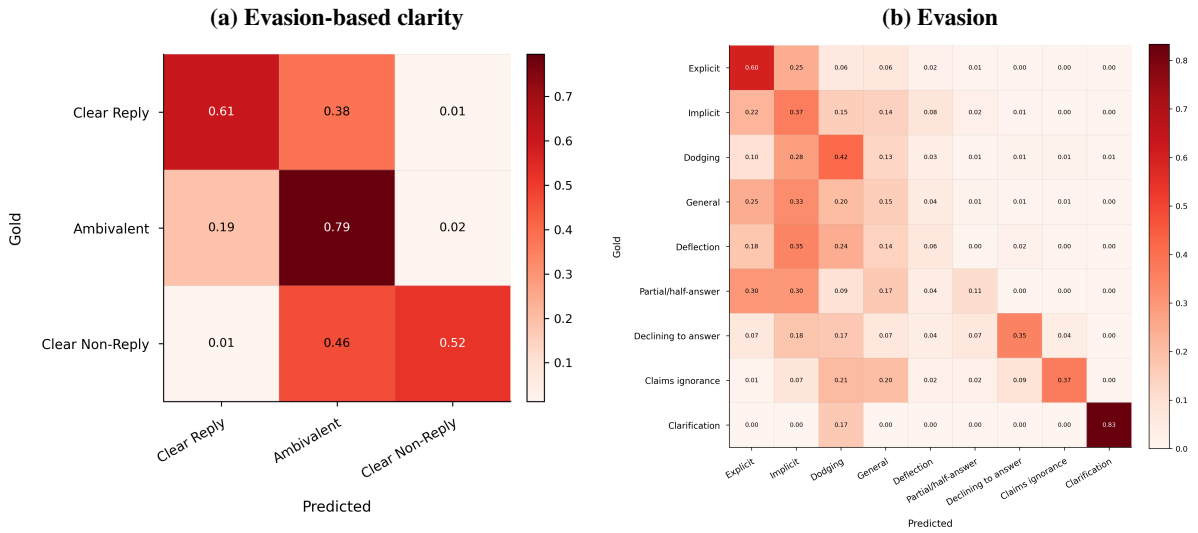

Figure 1: **Row-normalised confusion matrices on the test set for RoBERTa-large model.** Rows correspond to gold labels and columns to predicted labels; values represent row-wise proportions such that each row sums to 1. Raw counts are averaged across random seeds and, for the evasion task, across annotators prior to row normalisation.

(1) a majority-vote ensemble of five RoBERTa-large models trained on evasion labels, with clarity inferred via the taxonomy, and (2) zero-shot GPT-5.2 with the prompt in Appendix K. While encoder results in this paper are averaged over three seeds for scientific rigour, the submitted ensemble uses five seeds with per-instance majority voting.

Our best-performing system on the hidden set is GPT-5.2, achieving evasion-based clarity F1 = 0.74 (Task 1, ranked 22nd out of 44) and evasion F1 = 0.50 (Task 2, ranked 13th out of 33) (Thomas et al., 2026). The RoBERTa-large ensemble achieved lower scores, with F1 = 0.72 for evasion-based clarity and F1 = 0.45 for evasion.[5] This reverses the trend observed on the public test set (Section 5.1), where encoder-based models outperform GPT-5.2.

Notably, both systems improved substantially from public to hidden test (RoBERTa-large clarity: 0.661 → 0.72; GPT-5.2 clarity: 0.626 → 0.74). Given the tight variance reported in Table 3, we do not attribute the encoder improvement only to the five-seed ensemble. We hypothesise that this gap may reflect distributional differences between the public and hidden test sets, potentially favouring labels on which models performed better. Additionally, GPT-5.2 outperforming the encoders on the hidden set suggests that the fine-tuned models may have partially overfitted to the training distribution, though the marked improvement of encoders on the hidden set complicates this interpretation.

5.4 Error Analysis

Figure 1 shows row-normalised confusion matrices for both tasks for the RoBERTa-large model, with per-label breakdowns in Appendix F.

At the clarity level, the *Ambivalent Reply* class achieves the highest recall (0.79), followed by *Clear Reply* (0.61) and *Clear Non-Reply* (0.52). Most errors occur at the boundaries between *Ambivalent Reply* and the other two classes, consistent with the IAA patterns reported in Section 2.2. At the evasion level, the model shows strong recognition for *Clarification* (0.83) and *Explicit* (0.60), but exhibits substantial confusion among *Implicit*, *General*, *Deflection*, and *Partial/half-answer*. This mirrors known areas of annotator disagreement and likely reflects semantic overlap between fine-grained evasion strategies, suggesting that further gains may require modelling annotation uncertainty (rather than only additional model optimisation).

6 Conclusion

In this paper, we presented our contribution to the CLARITY shared task at SemEval 2026. We compared fine-tuned encoder models with zero-shot decoder-only models. Evasion-based clarity performed comparably to direct clarity. RoBERTa-large was strongest on the public test set, whereas zero-shot GPT-5.2 generalised better on the hid-

[5]While only the best system was recorded on the leaderboard, our RoBERTa-large ensemble would have ranked approximately 25th out of 44 for Task 1 and 18th out of 33 for Task 2 if it had been submitted instead of our best system (Thomas et al., 2026).

den evaluation set, suggesting a trade-off between in-domain performance and robustness. Most auxiliary training variations we attempted yielded no consistent improvements.

Overall, detecting political evasion remains challenging, as reflected by annotator disagreement (Section 2.2). Moreover, results on the test sets should be interpreted cautiously: several labels have low support, and macro F1 can be sensitive to a small number of predictions for rare classes (Table 11 and 12 in Appendix F).

Limitations

Conflation in input representation analysis. The ablation discussed in Section 3.3 and Appendix D compares *Segmented* and *Marked* input formats. However, this comparison simultaneously varies two factors: the ordering of the question–answer fields and the mechanism used to mark their boundary. As a result, the observed performance differences cannot be attributed unambiguously to either factor. A more controlled analysis that isolates field ordering from boundary representation would provide clearer insight into which component drives the improvement. We do not pursue this analysis here, as it falls outside the primary scope of the shared task and our focus in this work.

Annotation and supervision signals. A further limitation concerns the supervision signals used for both tasks. For clarity classification, training relies on a single aggregated label per instance. Given the subjective nature of the task, soft labels that reflect the distribution of annotator judgements could better capture annotation uncertainty and potentially provide richer supervision. For the evasion task, multi-annotator labels are available only at test time, whereas training uses a single label per instance. This mismatch may underestimate model performance and fails to reflect the overlapping nature of evasion strategies, where multiple interpretations may be plausible. A multi-label formulation or training with annotator distributions could therefore provide a more faithful modelling framework.

Zero-shot evaluation of decoder-only models. In this work, decoder-only models were evaluated only in a zero-shot setting, without any supervised fine-tuning on the QEvasion dataset. While this design allows us to assess the out-of-the-box capabilities and cross-domain robustness of large language models (LLMs), it does not reflect their full potential under task-specific training. In particular, parameter-efficient fine-tuning methods may enable these models to better capture the task taxonomy and domain-specific discourse patterns. Future work could therefore explore supervised fine-tuning of large decoder-only models, including parameter-efficient approaches such as LoRA (Hu et al., 2022), as well as cross-domain transfer learning to improve the generalisability of encoder-based models.

Performance gap to top systems. The performance gap between our systems and the top-ranked submissions highlights a methodological limitation of the approaches we explored. In particular, our strongest systems rely on either single-pass encoder-based classification or zero-shot prompting, whereas the task overview (Thomas et al., 2026) indicates that the best performing submissions typically benefited from hierarchical decomposition, multi-stage inference, confidence-based routing, and more elaborate prompt design. These differences are especially relevant for evasion-level classification, where several categories are semantically close and were also reported by the organisers to be difficult both for systems and annotators to distinguish. Our models use the task taxonomy during training and label mapping, but they do not explicitly exploit that hierarchy at inference time through staged decision-making or branch-restricted prediction. As a result, they are likely less effective at resolving fine-grained distinctions such as those between *General*, *Deflection*, and *Implicit*. However, higher benchmark performance from more complex pipelines does not necessarily indicate a more general solution: methods that are tightly coupled to the shared task taxonomy may achieve stronger benchmark performance while being less informative about transfer to other datasets or political settings. For this reason, we prioritised comparatively simple and reproducible methods, which we view as useful baselines even if they do not reach the performance of the most competitive task-specific systems.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/W524475/1]. The authors acknowledge the use of the Computational Research, Engineering and

Technology Environment (CREATE) at King's College London (King's College London, 2025). The authors also thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this work.

References

AI@Meta. 2024. Llama 3 model card.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Janet Beavin Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. 1988. Political equivocation: A situational explanation. *Journal of Language and Social psychology*, 7(2):137–145.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. *Equivocal communication.* Sage Publications, Inc.

Peter Bull. 1998. Equivocation theory and news interviews. *Journal of Language and Social Psychology*, 17(1):36–51.

Peter Bull. 2003. *The microanalysis of political communication: Claptrap and ambiguity*. Routledge.

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*.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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*.

King's College London. 2025. King's computational research, engineering and technology environment (create). https://doi.org/10.18742/rnvf-m076. Retrieved December, 2025.

Khaled Al Nuaimi, Gautier Marti, Alexis Marchal, and Andreas Henschel. 2025. Detecting evasive answers in financial Q&A: A psychological discourse taxonomy and lightweight baselines. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 191–196, Suzhou, China. Association for Computational Linguistics.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Preprint*, arXiv:2003.12206.

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

Archie Sage, Jeroen Keppens, and Helen Yannakoudakis. 2025. A survey of cognitive distortion detection and classification in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14884–14899, Suzhou, China. Association for Computational Linguistics.

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine learning-based research: Overview, barriers and drivers. *Preprint*, arXiv:2406.14325.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay

Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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. *Preprint*, arXiv:2409.13879.

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.

Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson. 2011. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company.

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.

Appendix

A QEvasion Dataset Analysis

As outlined in Section 2, the CLARITY task comprises two hierarchically related classification tasks: (Task 1) clarity level classification and (Task 2) evasion level classification. An illustrative example of a question–answer pair with its associated labels is presented in Figure 2. The definitions of the class labels for both levels are presented in Table 5.

Figure 3 presents the distribution of clarity and evasion level classes in the training set. The most frequent clarity label is *Ambivalent Reply* (59.2%). Within this category, *Dodging* is the most common evasion technique, whereas *Partial/half-answer* is the least frequent. The second most prevalent clarity label is *Clear Reply* (30.5%), while *Clear Non-Reply* (10.3%) occurs least often. Overall, the figure indicates a substantial class imbalance in both tasks.

A similar, although slightly different, distribution is observed in the public test set, where *Ambivalent Reply* accounts for 66.9% (206 instances), followed by *Clear Reply* with 25.6% (79 instances), and *Clear Non-Reply* with 7.5% (23 instances), confirming the persistence of class imbalance across data splits.

For the evasion level, the test set includes three labels per instance, each provided by a different annotator. In 275 out of 308 cases (89.3%), a final evasion label can be assigned through majority voting. However, in 33 instances (10.7%), no majority label can be determined, as all three annotators assigned different labels. While our evaluation procedure preserves disagreement (Section 4.2), the majority-vote breakdown above is useful for indicating how often a single-label collapse would require a tie-break.

Figure 4 shows the distribution of presidents alongside the clarity level distribution in the training set. The most represented president is *Trump* with 1,325 samples (38.4%), followed by *Obama* with 1,010 (29.3%), *Bush* with 714 (20.7%), and *Biden* with 399 (11.6%). This indicates that the dataset is also imbalanced with respect to presidential interviews.

Overall, the four presidents exhibit similar proportions of *Clear Reply*, *Ambivalent Reply*, and *Clear Non-Reply* instances. An exception is *Obama*, who shows a higher percentage of ambivalent replies and a comparatively lower percentage

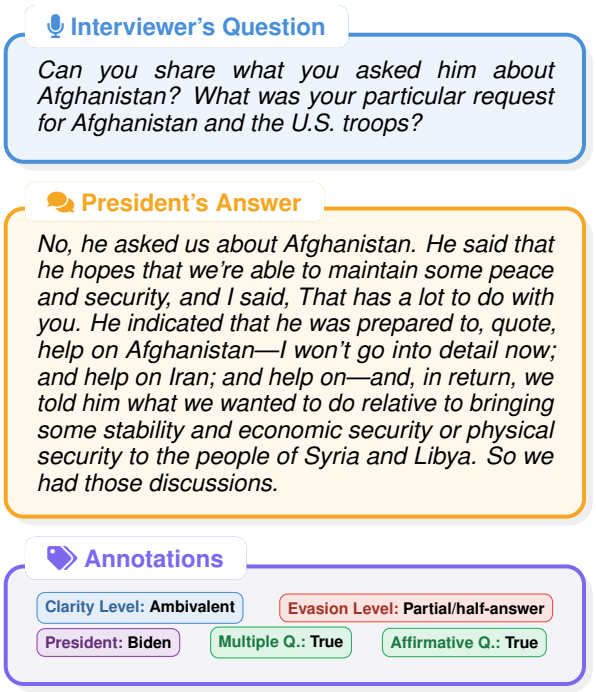

Figure 2: An example from the CLARITY dataset

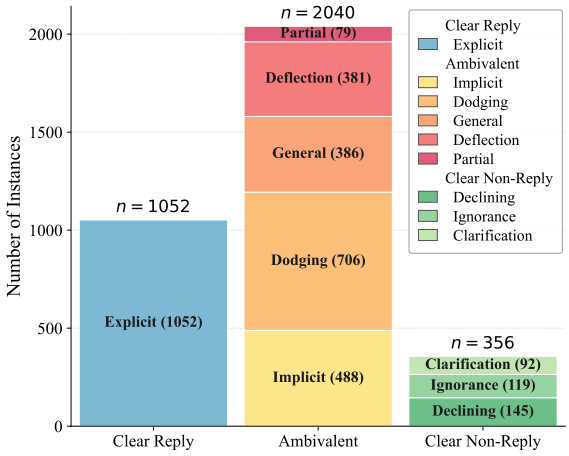

Figure 3: Class distribution in the training set.

of clear replies. President labels are not provided for the test set.

B Masking Ablation Study

In the QEvasion dataset paper, the authors explored a 'prior knowledge hypothesis', observing that clarity level classification performance was consistently higher on instances that did *not* contain named entities (Thomas et al., 2024). They attribute this effect to the fact that named entities often carry implicit, commonly assumed properties that are not explicitly stated in the response, therefore requiring models to rely on encoded world knowledge rather than surface-level discourse cues

Label	Definition
Clarity Level	
Clear Reply	Answers that admit a single interpretation.
Clear Non-Reply	Answers that openly refuse to share information.
Ambivalent Reply	Valid answers that allow for multiple interpretations.
Evasion Level	
Explicit	The requested information is clearly and directly stated in the expected form.
Implicit	The requested information is provided, but not explicitly stated or not presented in the expected form.
Dodging	The question is completely ignored.
General	A response is given, but it is overly general and lacks the specific details requested.
Deflection	The response initially addresses the topic but then shifts focus, making a different point than the one asked.
Partial	The response provides only part of the required information, addressing only a specific part of the request.
Declining	The question is acknowledged, but the respondent directly or indirectly refuses to answer.
Ignorance	The respondent explicitly states that they do not know the answer.
Clarification	The respondent does not provide the requested information and instead asks for clarification.

Table 5: Clarity and evasion technique label definitions, adapted from Thomas et al. (2024).

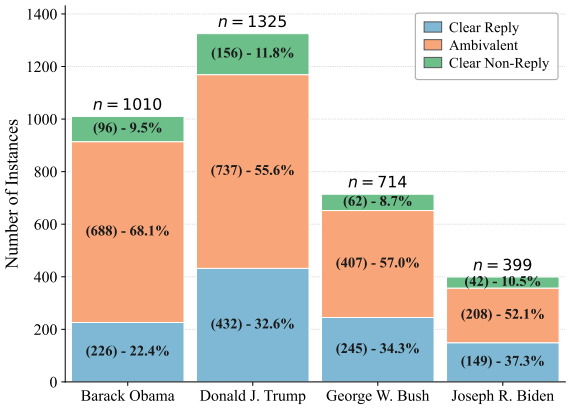

Figure 4: President and clarity level distribution in the training set.

alone. As this effect was reported to be most pronounced for smaller models, we conducted a series of experiments using RoBERTa-base, both to reduce computational cost and to assess whether this phenomenon would persist in a more controlled setting. This observation motivated our decision to experiment with different masking strategies: (i) no masking, (ii) naive masking, and (iii) entity-aware masking. All masking experiments were conducted at the clarity level.[6]

B.1 Variants

Naive masking. In the naive masking setting, person-name mentions are replaced with a single generic token `[PERSON]`. Person entities are identified using spaCy's named entity recogniser with the `en_core_web_lg` model, and each detected mention is masked independently, without preserving identity across mentions or between the question and the answer. An example of naive masking is shown below:

> **Question:** Did John Smith support the proposal?
>
> **Masked Question:** Did `[PERSON]` support the proposal?
>
> **Answer:** John said that Mary Johnson was responsible.
>
> **Masked Answer:** `[PERSON]` said that `[PERSON]` was responsible.

As all person mentions are mapped to the same placeholder, this approach removes explicit name information but collapses identity distinctions.

Entity-aware masking. Entity-aware masking extends naive masking by preserving consistency across references to the same individual within a question-answer pair. Person entities are again detected using spaCy's `en_core_web_lg` model, but distinct placeholders of the form `[PERSON_i]` are assigned to different individuals based on surface-form similarity. The same example under entity-aware masking becomes:

> **Question:** Did John Smith support the proposal?
>
> **Masked Question:** Did `[PERSON_1]` support the proposal?

[6] Masking implementation details are provided in the accompanying code.

Table 6: **RoBERTa-base performance under different masking strategies for direct clarity classification on the public test split.** We report macro-F1 (mean ± standard deviation over three seeds). Experiments use a different training setup from the main baselines, so scores are not directly comparable across tables; comparisons within this table remain valid.

Masking Strategy	Macro-F1
None	0.585 ± 0.013
Naive	0.570 ± 0.006
Entity-aware	0.566 ± 0.016

Answer: John said that Mary Johnson was responsible.

Masked Answer: `[PERSON_1]` said that `[PERSON_2]` was responsible.

The underlying idea was to encourage the model to rely on relative reference patterns between the question and the answer, rather than on memorised world-level knowledge associated with specific entities.

B.2 Results

Table 6 reports RoBERTa-base performance under the three masking settings, with all other training choices held constant and only the masking strategy varied. Across runs, the unmasked baseline attains the highest mean macro-F1 (0.585 ± 0.013), while both masking variants yield slightly lower mean scores on the HuggingFace test split (naive: 0.570 ± 0.006, entity-aware: 0.566 ± 0.016). Overall, these results suggest that, for this encoder model and training setup, masking person entities does not provide a clear benefit for clarity-level classification. Differences between the two masking variants are modest, and we do not observe a consistent advantage of entity-aware over naive masking in this setting.

However, these results should not be taken as direct evidence against the task paper's 'prior-knowledge hypothesis'. Our masking only replaces *person* names, and it depends on automatic entity detection and fairly simple matching, which can introduce noise and remove useful cues (for example, who is being referred to across the question and answer - especially for naive masking). Also, lower macro-F1 on this test split does not rule out the possibility that masking could help in other settings - for instance, when evaluating on interviews from different sources or speakers - which we leave for future work.

C Loss Weighting Ablation Study

As introduced in Section 3.2, we evaluate RoBERTa-large under three loss-weighting strategies: (i) *Unweighted*, (ii) *Balanced*, and (iii) *Sqrt*, to account for class imbalance.

Table 7 reports results on the public test set. Overall, weighting does not yield consistent improvements. The unweighted model performs best, achieving an evasion-based clarity macro F1 of 0.661. For evasion classification, it attains $\mathrm{ACC}_{\mathrm{match}} = 0.539$ and an average macro F1 of 0.371 across annotators. Both weighting schemes reduce performance on average and increase variance, particularly under the more aggressive balanced setting.

D Input Representation Ablation Study

Table 8 compares the two input formats described in Section 3.3 for RoBERTa-base. The marked representation (`[QUESTION]` q `[ANSWER]` a) consistently outperforms the segmented format (`[CLS]` a `[SEP]` q `[SEP]`) on both tasks, improving evasion-based clarity macro F1 from 0.518 to 0.595 and evasion $\mathrm{F1}_{\mathrm{avg}}$ from 0.270 to 0.338. The same trend holds for direct clarity prediction (Table 9), where the marked format improves macro F1 from 0.573 to 0.598. This suggests that explicit boundary tokens and/or presenting the question before the answer provides a stronger inductive bias for modelling question-answer alignment. As the creators of the QEvasion dataset note, limited context windows may cap encoder performance (512 tokens in this case for RoBERTa-base) (Thomas et al., 2024). One possible explanation is that placing the question before the answer reduces the likelihood that the question - the primary grounding signal - is truncated when inputs exceed the maximum length, even if this increases truncation of the answer. We leave a more controlled analysis of truncation effects (e.g., varying maximum length and truncation strategy) to future work.

E Stratified vs President-based Splitting Ablation Study

As described in Section 4.1, we evaluate both label-stratified and president-disjoint splitting strategies. In the label-stratified setting, we apply *dual stratification* over both clarity and evasion labels. In the president-disjoint setting, all responses from

Table 7: **RoBERTa-large loss-weighting ablation on the public test set.** For evasion-based clarity we report macro-F1 (F1), precision (P), and recall (R). For evasion we report $\mathrm{ACC}_{\mathrm{match}}$, the fraction of predictions matching at least one annotator, per-annotator macro-F1, and the average macro-F1 across annotators. All models are fine-tuned on evasion labels, with clarity inferred via the taxonomy. Results are averaged over three seeds (mean on the main row, standard deviation in brackets on the row beneath). Best-performing metrics are shown in bold.

Loss	*Evasion-based clarity*			*Evasion*				
	F1	**P**	**R**	$\mathbf{ACC_{match}}$	$\mathbf{F1_{A1}}$	$\mathbf{F1_{A2}}$	$\mathbf{F1_{A3}}$	$\mathbf{F1_{avg}}$
Unweighted	**0.661** (0.022)	**0.694** (0.013)	0.641 (0.023)	**0.539** (0.011)	**0.363** (0.030)	**0.378** (0.035)	**0.374** (0.047)	**0.371** (0.037)
Balanced	0.603 (0.077)	0.602 (0.098)	**0.642** (0.023)	0.515 (0.049)	0.350 (0.033)	0.347 (0.035)	0.327 (0.047)	0.341 (0.038)
Sqrt	0.602 (0.046)	0.617 (0.052)	0.625 (0.017)	0.505 (0.024)	0.318 (0.036)	0.304 (0.041)	0.302 (0.034)	0.308 (0.035)

Table 8: **RoBERTa-base input representation ablation on the public test set.** For evasion-based clarity we report macro-F1 (F1), precision (P), and recall (R). For evasion we report $\mathrm{ACC}_{\mathrm{match}}$, the fraction of predictions matching at least one annotator, per-annotator macro-F1, and the average macro-F1 across annotators. All models are fine-tuned on evasion labels, with clarity inferred via the taxonomy. Results are averaged over three seeds (mean on the main row, standard deviation in brackets on the row beneath). Best-performing metrics are shown in bold.

Input	*Evasion-based clarity*			*Evasion*				
	F1	**P**	**R**	$\mathbf{ACC_{match}}$	$\mathbf{F1_{A1}}$	$\mathbf{F1_{A2}}$	$\mathbf{F1_{A3}}$	$\mathbf{F1_{avg}}$
Segmented	0.518 (0.019)	0.520 (0.018)	0.534 (0.021)	0.473 (0.004)	0.277 (0.015)	0.272 (0.016)	0.261 (0.003)	0.270 (0.004)
Marked	**0.595** (0.028)	**0.589** (0.035)	**0.619** (0.030)	**0.516** (0.014)	**0.349** (0.024)	**0.332** (0.007)	**0.333** (0.008)	**0.338** (0.011)

Table 9: **RoBERTa-base input representation ablation for direct clarity on the public test set.** We report macro-F1 (F1), precision (P), and recall (R). Results are averaged over three seeds (mean on the main row, standard deviation in brackets on the row beneath). Best-performing metrics are shown in bold.

Input	*Direct clarity*		
	F1	**P**	**R**
Segmented	0.573 (0.010)	0.552 (0.018)	0.621 (0.020)
Marked	**0.598** (0.010)	**0.586** (0.013)	**0.636** (0.021)

a given president appear exclusively in one split, preventing speaker leakage.

Table 10 reports RoBERTa-large performance under the two strategies. The stratified split outperforms the president-disjoint split by approximately 0.04 in evasion-based clarity F1 and 0.05 in evasion $\mathrm{F1}_{\mathrm{avg}}$. This gap likely reflects the slight variation in label distributions across presidents noted in Appendix A. We note that the president-disjoint constraint applies only to the train/validation partition; the test set is shared across both configurations. The observed performance difference therefore primarily reflects the effect of a stricter model selection criterion rather than a direct measure of cross-speaker generalisation. A fully disjoint evaluation pipeline, including the test set, is left to future work.

F Per-label Performance Analysis

Table 11 reports precision, recall, and F1 per clarity label for the best-performing RoBERTa-large model. The *Ambivalent* class achieves the strongest results (F1 = 0.798), consistent with its larger support (206 instances). The *Clear Non-Reply* class shows high precision but low recall (0.522), suggesting the model tends to miss instances of this class, likely due to its limited support (23 instances). Finally, the *Clear Reply* class presents the weakest overall F1 (0.577), with relatively balanced but moderate precision and recall.

Table 12 reports the per-label support and F1 scores for each annotator separately and their average for the best-performing RoBERTa-large model. The *Clarification* class achieves the highest average F1 (0.841). However, its very low support (4 instances per annotator) limits the reliability of this estimate. *Explicit* is the strongest well-supported class (0.566), followed by *Declining to answer*

Table 10: **RoBERTa-large split-strategy ablation on the public test set.** For evasion-based clarity we report macro-F1 (F1), precision (P), and recall (R). For evasion we report $\mathrm{ACC}_{\mathrm{match}}$, the fraction of predictions matching at least one annotator, per-annotator macro-F1, and the average macro-F1 across annotators. All models are fine-tuned on evasion labels, with clarity inferred via the taxonomy. Results are averaged over three seeds (mean on the main row, standard deviation in brackets on the row beneath). Best-performing metrics are shown in bold.

Split Method	*Evasion-based clarity*			*Evasion*				
	F1	**P**	**R**	$\mathbf{ACC_{match}}$	$\mathbf{F1_{A1}}$	$\mathbf{F1_{A2}}$	$\mathbf{F1_{A3}}$	$\mathbf{F1_{avg}}$
Label-stratified	**0.661** (0.022)	**0.694** (0.013)	**0.641** (0.023)	**0.539** (0.011)	**0.363** (0.030)	**0.378** (0.035)	**0.374** (0.047)	**0.371** (0.037)
President disjoint	0.624 (0.029)	0.642 (0.027)	0.624 (0.035)	0.529 (0.012)	0.330 (0.036)	0.306 (0.033)	0.318 (0.058)	0.318 (0.039)

(0.424) and *Claims ignorance* (0.447). The most problematic classes are *Deflection* (0.084), *General* (0.184), and *Implicit* (0.277), all of which suffer from both low F1 and high variance across annotators, reflecting their inherent ambiguity and annotation inconsistency. All these classes belong to the ambivalent clarity level. Finally, *Partial/half-answer* is also poorly recognised (0.113), likely due to its extremely limited support across all annotators.

G Exploratory Experiments

G.1 Cross-Domain Transfer

As performance gains from training exclusively on the QEvasion dataset began to plateau during development, we explored the use of an additional dataset from a different domain, Earnings Calls Q&A[7] (Nuaimi et al., 2025). This dataset employs a Rasiah-style taxonomy (Rasiah, 2010), together with Bavelas forms (Bavelas et al., 1990) and Bull subtypes (Bull, 1998), maintaining continuity with established frameworks for analysing psychological and political equivocation. Importantly, its labelling scheme is derived from the same theoretical foundations as QEvasion, making Earnings Calls Q&A a strong candidate as a complementary resource. In particular, the Rasiah-style taxonomy uses the labels *Direct*, *Intermediate*, *Fully Evasive*, which align closely with QEvasion's clarity-level labels *Clear Reply*, *Ambivalent Reply*, and *Clear Non-Reply*, respectively.

Although the label sets are conceptually aligned, preliminary intermediate-task fine-tuning experiments were sensitive to optimisation settings and consistently reduced QEvasion validation macro-F1 in our setup (RoBERTa-large; Earnings Calls Q&A with Rasiah labels: 5 epochs; learning rate 3×10^{-5}; batch size 32; bf16; followed by continued fine-tuning on QEvasion). We did not investigate further whether this reflects domain mismatch or suboptimal training settings, and therefore excluded these results from the final system.

[7]Earnings Calls Q&A consists of question-answer exchanges between financial analysts and company executives during quarterly earnings announcements.

G.2 Cognitive Distortion Presence as an Auxiliary Signal

CDs are systematic patterns of biased reasoning studied in cognitive behavioural therapy (CBT). A growing body of work in NLP aims to detect them automatically in text (Sage et al., 2025). We hypothesised that distorted reasoning patterns in political responses - such as *Overgeneralisation* or *All-or-Nothing Thinking* - might provide a useful auxiliary signal for evasion detection, since evasive answers may exhibit superficially similar rhetorical patterns (e.g., vague generalisations or deflective framing).

To test this, we trained a lightweight CD detector using the CD dataset TherapistQA (Shreevastava and Foltz, 2021). We encoded texts with Sentence-BERT and fit a logistic regression model to predict binary distortion presence. We then scored each QEvasion training instance and discretised the resulting probabilities into two buckets (`CD_LOW` and `CD_HIGH`) using a threshold derived from the training distribution. These bucket tokens were prepended to the encoder input, allowing the model to condition on the estimated distortion level with minimal architectural changes.

In practice, this signal provided little discriminative value for either clarity or evasion classification. This may reflect a substantial domain gap between therapeutic dialogue and political interviews, where linguistic cues for CDs may not transfer cleanly. Additionally, the discretised token introduced noise and increased instability across

Table 11: **Per-label clarity (evasion-based) on the test set (RoBERTa-large).** Clarity labels are inferred from evasion via the taxonomy. Support is the number of test instances per label. Results are averaged over 3 seeds (13, 21, 42): mean on the main row, sample standard deviation in brackets on the row beneath.

Clarity label	Support	P	R	F1
Clear Reply	79	0.550 (0.005)	0.608 (0.046)	0.577 (0.019)
Ambivalent	206	0.801 (0.003)	0.794 (0.022)	0.798 (0.011)
Clear Non-Reply	23	0.733 (0.039)	0.522 (0.087)	0.608 (0.072)

Table 12: **Per-label evasion F1 on the test set (RoBERTa-large).** For each evasion label, we report F1 against each annotator (A1-A3) and the average across annotators. Support indicates the number of test instances for that label for each annotator. Results are averaged over 3 seeds (13, 21, 42): mean on the main row, sample standard deviation in brackets on the row beneath.

Evasion label	*Support*			*F1*			
	A1	A2	A3	$F1_{A1}$	$F1_{A2}$	$F1_{A3}$	$F1_{avg}$
Explicit	106	53	80	0.589 (0.033)	0.536 (0.027)	0.573 (0.028)	0.566 (0.030)
Implicit	54	54	67	0.242 (0.083)	0.257 (0.048)	0.333 (0.077)	0.277 (0.069)
Dodging	58	72	43	0.469 (0.015)	0.371 (0.035)	0.381 (0.029)	0.407 (0.021)
General	29	78	65	0.174 (0.019)	0.153 (0.050)	0.224 (0.055)	0.184 (0.029)
Deflection	30	22	23	0.062 (0.023)	0.093 (0.017)	0.098 (0.047)	0.084 (0.018)
Partial/half-answer	8	5	5	0.095 (0.086)	0.122 (0.113)	0.122 (0.113)	0.113 (0.104)
Declining to answer	10	9	14	0.383 (0.102)	0.519 (0.129)	0.370 (0.080)	0.424 (0.096)
Claims ignorance	9	11	7	0.411 (0.070)	0.508 (0.226)	0.422 (0.102)	0.447 (0.122)
Clarification	4	4	4	0.841 (0.167)	0.841 (0.167)	0.841 (0.167)	0.841 (0.167)

seeds. Given these results, we excluded this approach from our final systems. It is also important to note that discretising a scalar probability into two buckets is a blunt instrument and likely discards useful information (e.g., a continuous score or predicted CD type); we leave a more detailed investigation to future work.

H Hyperparameters and Training Setup

In Table 13, we report the hyperparameters and optimisation settings used across all encoder experiments. To ensure comparability, we fixed a single set of hyperparameters based on standard practice and applied them uniformly across all models, without any model-specific tuning. Full implementation details are available in our released codebase.

I Random Seeds

All encoder experiments are conducted using three fixed random seeds $(13, 21, 42)$. Results are reported as the mean across seeds, with the standard deviation shown in brackets in the tables. The random seed controls sources of stochasticity during training.

J Code, Compute, and Reproducibility

Code release and reproducibility. In line with recent calls for improved reproducibility in machine learning research (Semmelrock et al., 2025; Pineau et al., 2020), we release the full codebase upon publication to support reproducibility.

Compute. All encoder-based models reported in this paper were trained on a single NVIDIA A100-SXM4-40GB GPU. The total cumulative training

Table 13: **Training and optimisation settings for all encoder experiments.** All hyperparameters are shared across models except batch size and gradient accumulation steps, which are reduced for DeBERTa-v3-large to accommodate its larger memory footprint (effective batch size 32 throughout).

Setting	RoBERTa-base	RoBERTa-large	DeBERTa-v3-base	DeBERTa-v3-large
Max input length	512			
Learning rate	2×10^{-5}			
Warmup ratio	0.1			
Weight decay	0.01			
Dropout	0.1			
Max epochs	20			
Precision	bfloat16			
Batch size (train / eval)	32 / 32	32 / 32	32 / 32	16 / 16
Gradient accumulation	1	1	1	2
Checkpoint selection	Best macro F1 (validation)			
Early stopping	Patience 5, threshold 10^{-3}			
Seeds	(13, 21, 42)			

runtime across all encoder experiments and multi-seed runs reported in this paper was 22,049 seconds (approximately 6.1 hours). For zero-shot experiments, we queried decoder-only models via hosted inference APIs (including the OpenAI API[8] for GPT5.2 and the Hugging Face Inference API[9] for open-weight models) and did not perform any additional fine-tuning.

K LLM Prompt

Figure 5 shows the system prompt used for zero-shot evasion classification. The prompt instructs the model to assign exactly one of the nine evasion labels to each input question-answer pair, following a structured decision ladder with tie-breaking rules and in-context mini-examples. To improve throughput, inference was performed in batches, with the model receiving multiple question-answer pairs in a single API call and returning predictions as a JSON array. The evasion-based clarity label was then derived deterministically from the predicted evasion label via the taxonomy.

[8] https://developers.openai.com/api/docs/

[9] https://huggingface.co/docs/inference-providers/en/index

```
system_prompt: |
  You are an expert annotator for CLARITY Task B (evasion level).
  Input: a QUESTION and an ANSWER (political interview style).
  Output: EXACTLY ONE label for each item from this set:
  Explicit, Implicit, Dodging, General, Deflection,
  Partial/half-answer, Declining to answer, Claims ignorance, Clarification.

  Core principle: decide based on whether the ANSWER supplies the *requested commitment*.
  Requested commitment = the specific yes/no, person, time, place, number, policy stance,
  or concrete plan the QUESTION asks for.
  If that commitment is present (even indirectly), it is NOT Dodging/Deflection.

  Step 0 - normalise the question:
  - Treat multi-part questions as requiring ALL parts unless the question clearly
    foregrounds one part.
  - If the question contains a yes/no + 'why/how/what specifics', then a bare
    yes/no is incomplete.

  Decision ladder (apply in order; stop at the first that clearly applies):
  1) Clarification
     - The answer primarily asks the interviewer to repeat/clarify/restate.
     - If it both asks for clarification AND gives the requested commitment,
       choose based on the dominant function.

  2) Claims ignorance
     - The answer asserts lack of knowledge/recall/awareness.
     - IMPORTANT: If the answer later gives the requested commitment anyway,
       do NOT pick Claims ignorance.

  3) Declining to answer
     - The answer refuses to provide the requested commitment now.
     - If it refuses but then gives the requested commitment, label by what dominates.

  4) Explicit
     - Directly states the requested commitment (even if brief), in the form requested.

  5) Implicit
     - The requested commitment is not stated verbatim, but is clearly recoverable
       via a straightforward inference.
     - Test: a reasonable listener could paraphrase the commitment in one sentence
       without guessing.

  6) Partial/half-answer
     - Answers ONE required part but omits other required parts.

  7) General vs Deflection vs Dodging (non-answer family)
     7a) General    - On-topic, but avoids the requested commitment by staying
                      vague or non-committal.
     7b) Deflection - Acknowledges the question then shifts to a different frame.
                      No requested commitment.
     7c) Dodging    - Does NOT engage the requested commitment AND does NOT
                      meaningfully address the topic.

  Tie-breakers:
  - Concrete list of steps in response to 'what will you do'
    -> prefer Explicit (or Partial if incomplete).
  - Gives constraints but never the asked commitment -> Declining to answer.
  - Explains the topic but not the specific asked stance -> General or Deflection,
    not Dodging.
  - 'I already answered that' without the commitment -> Declining to answer.

  Mini-examples:
  A) Dodging vs Deflection
  Q: Did you meet the leader yesterday?
  A (Dodging):     Our country is doing very well economically.
  A (Deflection):  I understand the question - what matters is focusing on jobs
                   for families.

  B) General vs Implicit
  Q: Do you support the bill?
  A (General):  We're considering all options and working with colleagues.
  A (Implicit): I'll be voting yes when it reaches the floor.

  C) Declining vs General
  Q: What concessions do you want them to make?
  A (Declining): I'll tell you after the meeting.
  A (General):   We want cooperation and stability.

  D) Claims ignorance vs General
  Q: Have you seen evidence of acceleration?
  A (Claims ignorance): I haven't seen that report.
  A (General):          We're concerned and monitoring closely.

  Output rules:
  - Choose exactly one label per item.
  - Keep the item order exactly.
  - Judge using only the provided QUESTION and ANSWER.

  Respond ONLY with a JSON object in this exact format: {"labels": ["label1", "label2", ...]}
  No explanation, no markdown fences, only the JSON.
```

Figure 5: Zero-shot prompt for evasion label prediction and evasion-based clarity.