ttda704 at SemEval-2026 Task 6: Structured Chain-of-Thought Prompting for Political Evasion Detection

Tai Tran Tan[1,2*], **An Dinh Thien**[1,2*]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{22521287, 22520010}@gm.uit.edu.vn

Abstract

This paper describes our system[1] for SemEval-2026 Task 6, which addresses the classification of political evasion strategies in English question–answer pairs extracted from U.S. presidential interviews. We systematically compare two distinct paradigms: (1) Parameter-Efficient Fine-Tuning of Qwen3 models (4B–32B) using QLoRA, enhanced with tiered upsampling and weighted cross-entropy loss to address severe class imbalance, and (2) structured Chain-of-Thought (CoT) prompting of reasoning-capable API models, namely DeepSeek-V3.2 and Grok-4-Fast. Our evaluation demonstrates that structured CoT prompting of reasoning-enabled models substantially outperforms our baseline parameter-efficient fine-tuning implementation in absolute Macro F1. Our best system, Grok-4-Fast with extended reasoning and few-shot hierarchical CoT prompting, achieves a Macro F1 of 0.5147 on Subtask 2 (9-class evasion) and 0.7979 on Subtask 1 (3-class clarity), ranking 8/33 on Subtask 2 and 13/41 on Subtask 1 on the official leaderboard. Furthermore, our ablation studies reveal key insights into effective prompt design for evasion detection: presenting labels within a hierarchical taxonomy helps structure model reasoning while few-shot exemplars provide task calibration; however, the strongest prompt variants are not statistically distinguishable in Macro F1, and explicitly enabling extended reasoning modes yields substantial performance gains by facilitating the multi-step pragmatic analysis required to detect evasive intent.

1 Introduction

Political discourse is strategically ambiguous. In televised interviews, politicians routinely employ evasion techniques such as topic shifts, deflections, and refusals that undermine democratic transparency. Prior work reports that politicians provide clear answers to only 39–46% of questions, compared to 70–89% for non-politicians (Bull, 2003). Automatically detecting these strategies is challenging because evasion often relies on pragmatic inference rather than surface-level cues.

SemEval-2026 Task 6 (Thomas et al., 2024), *CLARITY: Unmasking Political Question Evasions*, focuses on classifying semantic evasion in political question-answer (QA) pairs. The task poses two evaluation challenges: **Subtask 1 (Clarity)** categorizes responses into three coarse tiers (*Clear Reply*, *Ambivalent*, *Clear Non-Reply*), while **Subtask 2 (Evasion)** demands fine-grained classification into nine specific evasion strategies (e.g., *Dodging*, *Deflection*, *Implicit*).

Standard text classifiers handle overt evasion but struggle with subtle, ambivalent replies because they lack explicit pragmatic reasoning. Separating a *General* platitude from an *Implicit* answer, or a tangential *Deflection* from full *Dodging*, requires context-sensitive inference beyond surface matching.

We address this by systematically studying political evasion classification via two paradigms: PEFT of open-weight models (Qwen3 4B–32B) and structured CoT prompting with reasoning-focused APIs (DeepSeek-V3.2, Grok-4-Fast). To mitigate the difficulty of the flat 9-class taxonomy, we introduce hierarchical prompts that evaluate "Directness" and "Topic Fidelity" before the final label (Figure 1).

Our main contributions are as follows:

1. We identify that the lack of intermediate pragmatic reasoning is a core bottleneck for classifying ambivalent evasion. To overcome this, we introduce step-by-step CoT prompting templates that explicitly condition models to evaluate linguistic directness and topic alignment.

2. We show that restructuring a flat class taxonomy into a conceptual hierarchy provides a

[*]Equal contributions.
[1]Code and prompts are available at `https://github.com/taitran501/SemEval-2026-Task6`

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2817–2829
July 3-4, 2026 ©2026 Association for Computational Linguistics

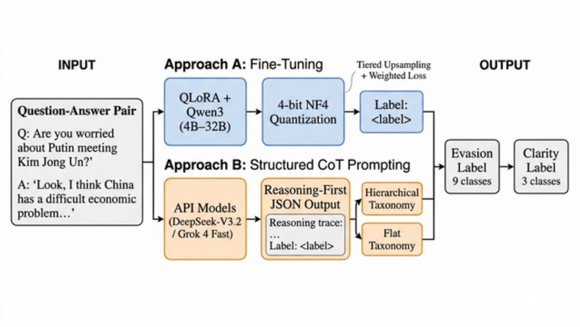

Figure 1: Overview of our two pipelines: Approach A (QLoRA fine-tuning) and Approach B (structured CoT prompting).

useful reasoning scaffold for distinguishing closely related evasion boundaries, helping capable models organize fine-grained labels and produce clearer reasoning traces.

3. We analyze model errors and show that CoT reasoning can over-interpret vague political language, especially around *Dodging*, *Deflection*, *General*, and *Implicit*.

2 Related Work

Political evasion has long been studied in discourse and political communication, where prior work documents how public figures avoid, partially answer, or reframe questions in interviews (Bull, 2003; Bull and Mayer, 1993; Clayman, 2001; Rasiah, 2010). Recent NLP work operationalizes this phenomenon through response-clarity and evasion taxonomies, including the QEvasion dataset and the CLARITY shared task (Thomas et al., 2024; Ferracane et al., 2021).

Our work relates to two lines of NLP research. First, hierarchical label information can help models reason over related political labels (Dayanik et al., 2022). Second, chain-of-thought and structured prompting can elicit multi-step reasoning in LLMs (Wei et al., 2022; Wang et al., 2023), while PEFT methods such as LoRA and QLoRA enable efficient adaptation of open-weight models (Hu et al., 2022; Dettmers et al., 2023). We compare these two paradigms under a controlled shared-task setting for political evasion detection.

3 Task Description

3.1 Task Definition

SemEval-2026 Task 6 (Thomas et al., 2024), *CLARITY: Unmasking Political Question Evasions*, frames the detection of political evasion as a classification task over question–answer (QA) pairs drawn from U.S. presidential interviews spanning 2006–2023. Multi-part interview questions are first decomposed into singular sub-questions, so that each annotation precisely captures how well a single specific inquiry is addressed. Systems are evaluated on two nested subtasks:

- **Subtask 1 (Clarity):** Classify the respondent's answer into one of three coarse clarity tiers: *Clear Reply*, *Ambivalent*, or *Clear Non-Reply*.

- **Subtask 2 (Evasion):** Classify the answer into one of nine fine-grained evasion strategies. The clarity label is deterministically derived from the evasion label via a fixed mapping.

The nine evasion labels and their parent clarity categories are summarized in Table 1.

Clarity (Subtask 1)	Evasion (Subtask 2)
Clear Reply	Explicit
Ambivalent	Implicit General Dodging Deflection Partial/half-answer
Clear Non-Reply	Declining to answer Claims ignorance Clarification

Table 1: The two-level taxonomy of response clarity. Each evasion label (Subtask 2) maps to exactly one clarity category (Subtask 1).

3.2 Evaluation

Both subtasks are evaluated using macro-averaged F1-score, ensuring equal weight across all classes regardless of their frequency. Evaluation is conducted on the official test set released by the organizers, as well as a held-out private evaluation set used for the final competition leaderboard.

3.3 Dataset

The dataset (Thomas et al., 2024) contains 3,448 training samples and 308 test samples. The training set is heavily imbalanced: *Explicit* (30.5%) and *Dodging* (20.5%) are the dominant classes, while *Partial/half-answer* (2.3%) and *Clarification* (2.7%) are severely underrepresented. Each instance provides three text fields: the full original interview question (`interview_question`), a decomposed sub-question (`question`), and the respondent's full answer (`interview_answer`).

4 Methodology

4.1 System Overview

We explore two complementary approaches: **Approach A** (Parameter-Efficient Fine-Tuning of local models) and **Approach B** (structured Chain-of-Thought prompting of API models). Both approaches emphasize explicit reasoning as a central mechanism for evasion detection (Wei et al., 2022; Wang et al., 2023). Figure 1 summarizes the two pipelines.

4.2 Approach A: Parameter-Efficient Fine-Tuning

We fine-tune the Qwen3 family (Yang et al., 2025) using QLoRA with 4-bit NF4 quantization (Dettmers et al., 2023). To counteract severe class imbalance, we apply $2\times$ upsampling to the three rarest evasion classes and apply inverse-frequency weighting (clipped at 5.0) to the cross-entropy loss. We evaluate four variants: Qwen3-4B, Qwen3-4B-Instruct-2507, Qwen3-14B, and Qwen3-32B.

Training configuration. Models are trained for a single epoch using Unsloth (Han et al., 2024) with learning rate 2×10^{-5} and a cosine scheduler. LoRA rank is set to 32 (alpha 64) across all linear projection layers. We proactively limit training to one epoch to minimize the risk of overfitting on the relatively small training set. Full per-model hyperparameters are detailed in Appendix A.

4.3 Approach B: Structured Chain-of-Thought Prompting

For API models, we enforce a reasoning-first output format consistent with prior CoT work (Wei et al., 2022). We evaluate two state-of-the-art architectures: DeepSeek-V3.2 (`deepseek-chat` vs. `deepseek-reasoner`) and Grok-4-Fast (`grok-4-fast` vs. `grok-4-fast-reasoning`). For both providers, we compare standard inference against extended reasoning modes that generate an internal chain-of-thought before the final answer. All models strictly output JSON containing a `reasoning` field followed by the `evasion_label` (Li et al., 2025).

System	S2 F1	S1 F1
Best Qwen3 PEFT	0.3630	0.5654
DeepSeek Rsn. Few-shot Hier.	0.5115	0.7735
Grok Rsn. Zero-shot Flat	0.5126	0.7821
Grok Rsn. Few-shot Hier.	**0.5147**	**0.7979**
Official submission†	0.5600	0.7900
Organizer revised baseline†	0.5700	0.8200

Table 2: Key results. S2 and S1 denote Macro F1 for Subtask 2 (evasion) and Subtask 1 (clarity), respectively. The top four rows are results on the 308-sample official test split with majority-vote labels. Rows marked with † are blind/private leaderboard scores and are shown only for context. The full 32-configuration comparison is in Appendix B.1.

4.4 Prompt Design

We provide the full `interview_question` as context and the decomposed `question` as the target, following Thomas et al. (2024). Prompts vary by in-context learning (zero-shot vs. few-shot: 9 exemplars for API CoT, 6 for fine-tuned standard) and taxonomy structure (hierarchical vs. flat).

Fine-tuned models output either standard text ('*Label: <label>*') or structured JSON (CoT). Together with the prompt axes, this yields four configurations per model. Prompt templates are detailed in Appendix C.

5 Results and Discussion

5.1 Experimental Setup

We evaluate all models on the 308-sample official test split. We report results using majority-vote gold labels from three independent annotators. Macro F1 is the primary metric for both subtasks. Subtask 1 F1 is derived by deterministically mapping predicted evasion labels to their parent clarity categories. For fine-tuned models with CoT outputs, labels are extracted using a robust three-stage pipeline of regex matching and string fallback (Appendix A).

5.2 Overview of Results

Table 2 summarizes the key systems; the full 32-configuration matrix is provided in Appendix B.1.

The best overall configuration on the 308-sample official test split is **Grok-4-Fast Reasoning, Few-shot CoT (hier.)**, achieving Macro F1 of **0.5147** on Subtask 2 and **0.7979** on Subtask 1. The best fine-tuned model (Qwen3-32B, Few-shot CoT) reaches 0.3630 on Subtask 2, so structured CoT prompting with reasoning-capable API models substantially outperforms our *baseline* PEFT setup. We interpret this comparison conservatively because it is confounded by model scale, disabled Qwen thinking mode, a single fine-tuning epoch, and different exemplar counts.

Official leaderboard. Our best submission (*ttda704*) ranks 8/33 on Subtask 2 (Evasion F1 = 0.56) and 13/41 on Subtask 1 (Clarity F1 = 0.79) on the official SemEval-2026 Task 6 leaderboard. The revised organizer baseline reported in the camera-ready task paper (Thomas et al., 2026) is 0.57 on Subtask 2 and 0.82 on Subtask 1.

Subtask 1 behavior. Subtask 1 scores are consistently higher because the coarse clarity taxonomy collapses several difficult ambivalent evasion labels into a single parent class. Errors such as *General* vs. *Implicit* or *Dodging* vs. *Deflection* are severe under the nine-way Subtask 2 metric, but often remain within the *Ambivalent* category and therefore do not hurt Subtask 1. This explains why models achieve relatively strong clarity F1 while still struggling with fine-grained evasion distinctions.

5.3 Ablation Analysis

Reasoning mode. Enabling extended reasoning consistently improves both API model families, with the largest gains under few-shot CoT. The benefit is especially clear for Grok, where reasoning plus zero-shot flat prompting already reaches 0.5126 Macro F1, nearly matching the strongest few-shot configuration.

Prompt structure and exemplars. The hierarchical taxonomy is best interpreted as a reasoning scaffold rather than a universally superior prompt format. It encourages models to first decide coarse clarity before selecting a fine-grained evasion label, which appears most useful for reasoning-capable models. Few-shot exemplars further calibrate ambiguous boundaries such as *General* vs. *Implicit* and *Dodging* vs. *Deflection*. However, the effect is not uniform: Grok-4-Fast Reasoning with zero-shot flat prompting reaches 0.5126 Macro F1, nearly matching the best few-shot hierarchical configuration at 0.5147.

Statistical significance. We assess the top configurations using approximate randomization tests (Yeh, 2000) (R = 10,000), McNemar's test (McNemar, 1947), and paired bootstrap confidence intervals (B = 10,000). The top three systems have heavily overlapping 95% confidence intervals: Grok Few-shot Hier. $[0.443, 0.565]$, Grok Zero-shot Flat $[0.436, 0.567]$, and DeepSeek Few-shot Hier. $[0.428, 0.569]$. Approximate randomization tests show no significant Macro F1 differences ($p > 0.19$), so the best configuration should be interpreted as the highest point estimate rather than a statistically superior system. McNemar's test nevertheless shows significant sample-level disagreement between Grok and DeepSeek few-shot hierarchical systems ($p = 0.015$), suggesting complementary error patterns.

Fine-tuned models. For local Qwen models, CoT prompting is not uniformly helpful, especially at 4B scale. The combination of simplified reasoning targets and output-format fragility means these results should be interpreted as a PEFT baseline rather than an upper bound.

Evasion Label	Prec	Rec	F1
Claims ignorance	0.7778	0.8750	0.8235
Clarification	1.0000	1.0000	1.0000
Declining to answer	0.6923	0.7500	0.7200
Deflection	0.2200	0.5238	0.3099
Dodging	0.8333	0.1852	0.3030
Explicit	0.6636	0.8022	0.7264
General	0.3571	0.2885	0.3191
Implicit	0.3103	0.3000	0.3051
Partial/half-answer	0.1000	0.1667	0.1250

Table 3: Per-class Subtask 2 performance for the best system (Grok-4-Fast Reasoning, Few-shot CoT) on the test set (N = 308).

5.4 Error Analysis

Table 3 reports per-class performance for the best system. The full confusion matrix and qualitative case studies are provided in Appendix E.

Performance drops sharply on pragmatically subtle categories. We observe three recurrent failure modes:

Dodging vs. other ambivalent forms. *Dodging* has critically low recall (0.1852), indicating that the model rarely identifies complete topic abandonment. When a true *Dodging* response is misclassified, the model predicts *Implicit* 26%, *General* 22%, and *Deflection* 20% of the time. A concrete mitigation is a second-stage topic-abandonment verifier that checks whether the answer preserves the core entity and requested information from the target question.

General vs. Implicit vs. Deflection. The model exhibits high mutual confusion across these ambivalent categories. Of true *General* errors, 29% are predicted as *Implicit* and 27% as *Deflection*, suggesting that step-by-step reasoning can over-extract specific intent from vague political language.

Over-prediction of Explicit. The model frequently defaults to the majority class. Notably, 42% of true *Implicit* responses are misclassified as *Explicit*, suggesting that topical relevance is often treated as sufficient evidence of directness. Appendix E.1.3 also shows that some apparent errors reflect genuine annotator disagreement rather than clear model failure.

5.5 Thinking Mode Follow-up (Qwen3-32B)

To address the confound that Qwen3's native thinking mode was disabled in the main PEFT experiments, we ran a diagnostic follow-up on a Qwen3-32B checkpoint. Because thinking-mode inference was substantially slower (approximately 100 minutes for 100 samples in our setup), we evaluated it on stratified samples of 100 instances rather than rerunning the full test set or full configuration grid. Enabling thinking improved Evasion F1 from 0.0734 to 0.2362 on the official-test sample setting and reduced hard fallback from 37.99% to 0.00%. On the blind sample, thinking reached 0.3700 Evasion F1 with a 1.00% hard-fallback rate. These results suggest that the original PEFT baseline is conservative, although this experiment is diagnostic rather than fully comparable to the main benchmark. Full details are reported in Appendix B.2 and Appendix A.5.

6 Conclusion

We compared QLoRA-based PEFT and structured CoT prompting for SemEval-2026 Task 6. Reasoning-capable API models substantially outperform our baseline PEFT setup in absolute Macro F1, although the comparison is confounded by model scale, native reasoning support, and exemplar count. Our best configuration, Grok-4-Fast Reasoning with few-shot hierarchical CoT, achieves the highest point estimate (0.5147 on Subtask 2 and 0.7979 on Subtask 1), but is not statistically superior to the strongest zero-shot flat variant. Overall, our findings suggest that hierarchical prompts are useful as reasoning scaffolds, few-shot examples act as task calibration, and the hardest remaining errors arise from pragmatic ambiguity around ambivalent evasion labels.

Limitations

Our PEFT results should be interpreted as a baseline rather than an upper bound: models were trained for one epoch without extensive hyperparameter search, and Qwen3 thinking mode was disabled in the main experiments. Our few-shot examples were manually curated and not systematically varied, which may introduce exemplar and speaker-style bias. Future work should evaluate exemplar sensitivity by sampling multiple few-shot sets across different presidents and by retrieving semantically similar exemplars at inference time; retrieval-based few-shot prompting may reduce speaker-style bias and improve coverage of rare evasion labels. We report blind/private leaderboard scores only for the submitted system rather than rerunning all 32 configurations, because full blind-set reruns would require substantial additional API calls and local inference time; controlled ablations are therefore conducted on the 308-sample official test split. Finally, the best result relies on Grok-4-Fast, a proprietary API model; the strongest open-weight reproducible configuration is DeepSeek-V3.2 Reasoning Few-shot CoT (hier.) at Macro F1 = 0.5115 on Subtask 2.

References

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

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14(4):651–666.

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

Erenay Dayanik, André Blessing, Nico Blokker, Sebastian Haunss, and Jonas Kuhn. 2022. Improving neural political statement classification with class hierarchical information. In *Findings of ACL*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1272–1286, Online. Association for Computational Linguistics.

Daniel Han et al. 2024. Unsloth: Fast and memory-efficient llm fine-tuning.

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 (ICLR)*.

Binghui Li, Minghui Zou, Xiaowang Zhang, Shizhan Chen, and Zhiyong Feng. 2025. A reinforcement learning framework for cross-lingual stance detection using chain-of-thought alignment. In *Findings of ACL*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

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

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.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, et al. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *ACL*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 947–953. Association for Computational Linguistics.

A Detailed Experimental Setup

A.1 Fine-Tuning Hyperparameters

Table 4 reports the full hyperparameter configuration for each fine-tuned model. The specific model checkpoints used are: `unsloth/Qwen3-4B` (4B base), `unsloth/Qwen3-4B-Instruct-2507-unsloth-bnb-4bit` (4B-Instruct), `unsloth/Qwen3-14B-unsloth-bnb-4bit` (14B), and `unsloth/Qwen3-32B-unsloth-bnb-4bit` (32B). All models share common LoRA parameters (rank 32, alpha 64, dropout 0.0) applied to all linear layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). Training uses AdamW 8-bit optimizer with cosine learning rate scheduling. Qwen3 thinking mode is disabled (`enable_thinking=False`) during both training and inference for all models.

Parameter	4B	4B-Inst	14B	32B
Quantization	NF4 (4-bit)			
LoRA rank / alpha	32 / 64			
LoRA dropout	0.0			
Learning rate	2×10^{-5}			
LR scheduler	Cosine			
Warmup ratio	0.05			
Epochs	1			
Optimizer	AdamW 8-bit			
Max class weight	5.0			
Batch / GPU	16	16	2	2
Grad. accum.	2	2	8	8
Eff. batch size	32	32	16	16
Weight decay	0.01	0.01	0.05	0.05
Max grad norm	1.0	1.0	0.5	0.5
Max seq. length	8092	8092	7000	7000
Max new tokens	1024	1024	2048	2048

Table 4: Full hyperparameter configuration for fine-tuned models. All models are trained with Unsloth for memory-efficient QLoRA.

A.2 Dataset Statistics and Inter-Annotator Agreement

Table 5 presents the complete distribution of evasion labels across training and test sets, revealing severe class imbalance with a 13.32:1 ratio between the most frequent (*Explicit*, 1052 samples) and least frequent (*Partial/half-answer*, 79 samples) classes in the training set. Figure 2 visualizes this distribution, highlighting that the three rarest classes (*Clarification*, *Claims ignorance*, *Partial/half-answer*) collectively represent only 8.41% of the training data.

Inter-Annotator Agreement. The dataset includes annotations from three independent annota-

Label	Train	Train %	Test	Test %
Explicit	1052	30.51	91	29.55
Implicit	488	14.15	60	19.48
General	386	11.19	52	16.88
Dodging	706	20.48	54	17.53
Deflection	381	11.05	21	6.82
Partial/half-ans.	79	2.29	6	1.95
Declining to ans.	145	4.21	12	3.90
Claims ignorance	119	3.45	8	2.60
Clarification	92	2.67	4	1.30
Total	3448	100.00	308	100.00

Table 5: Distribution of evasion labels in training and test sets. The class imbalance ratio is 13.32:1 (Explicit to Partial/half-answer).

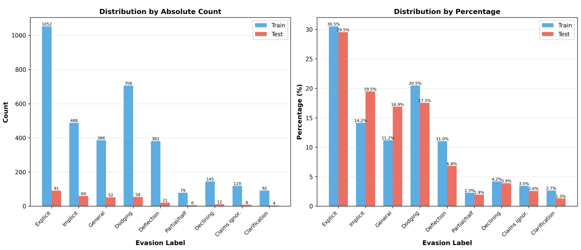

Figure 2: Visualization of label distribution across training and test sets, showing severe class imbalance particularly for the three rarest categories.

tors per instance. For the test set (N = 308), we observe full agreement (3/3 annotators) in 40.58% of cases, partial agreement (2/3 annotators) in 48.70%, and complete disagreement (all three different) in 10.71%. We calculate Fleiss' Kappa at 0.4723, indicating *moderate agreement*. This level of agreement is expected given the inherently subjective nature of pragmatic interpretation - human judges may legitimately focus on different aspects of the same response when applying the taxonomy.

The most common disagreement patterns reveal systematic ambiguities in the taxonomy boundaries: *General* vs. *Implicit* (24 cases), *Explicit* vs. *Implicit* (23 cases), and *Deflection* vs. *Dodging* (16 cases). These patterns directly correspond to the confusion clusters observed in our model predictions (Section 5.4), suggesting that even human annotators struggle with the same pragmatic boundaries that confound automated systems. Overall, 183 instances (4.87% of the full dataset) exhibit annotator disagreement, representing cases where political discourse interpretation is genuinely ambiguous rather than algorithmically deterministic.

A.3 Class Imbalance Handling

We address severe class imbalance via two complementary techniques:

1. **Tiered upsampling (2×)** of the three rarest classes: *Clarification* (92 → 184), *Claims ignorance* (119 → 238), and *Partial/half-answer* (79 → 158).

2. **Inverse-frequency class weights** computed as $w_c = \frac{N}{C \cdot n_c}$ (where N = total samples, C = number of classes, n_c = count of class c), clipped at 5.0 to avoid numerical instability.

A.4 Label Extraction Pipeline

For fine-tuned models generating CoT outputs, we extract evasion labels with a three-stage pipeline:

1. Regex match for `"evasion_label": "<label>"` in JSON output.

2. Fallback: extract text after `Label:` prefix.

3. Final fallback: string-match any valid label in the full output. If no label is found, default to *Explicit* (most frequent class).

This fallback is most frequently triggered for Qwen3-4B with CoT output format, where JSON parsing failures occur due to the model's limited instruction-following at small scale. The fallback may inflate Explicit-class precision and recall for these configurations.

A.5 Hard-Fallback Statistics

Table 6 reports the hard-fallback rate for the Qwen3-32B model under different inference configurations. A hard fallback (defaulting to *Explicit*) indicates a complete output format failure rather than a genuine prediction, and its rate therefore directly reflects the reliability of the extraction pipeline for a given configuration.

The near-complete elimination of hard fallbacks when thinking is enabled suggests two effects: (1) the model's extended reasoning process produces more structured, parseable outputs; and (2) the non-thinking hard-fallback rate artificially inflates *Explicit*-class predictions in the main results table, making the non-thinking baseline appear worse than a configuration with stable output formatting.

We report the Qwen3-32B diagnostic setting because it directly affects the thinking-mode follow-up in Section 5. A full per-configuration fallback

Model	Inference Mode	Split	N	Hard Fallback
Qwen3-32B	Non-thinking	Official Test	308	37.99% (117/308)
Qwen3-32B	Thinking	Official Test	100	**0.00%** (0/100)
Qwen3-32B	Thinking	Blind Set	100	**1.00%** (1/100)

Table 6: Hard-fallback rates for Qwen3-32B. The high rate in non-thinking mode (37.99%) indicates systematic output format failure; enabling thinking inference reduces this to near zero, demonstrating that the reasoning mode substantially stabilises output formatting in addition to improving classification quality.

audit would require reprocessing all raw generations and is outside the scope of this camera-ready revision; however, structured CoT outputs are the most affected by complete parsing failure, whereas standard label-only prompting is less prone to hard fallback.

A.6 API Model Configuration

DeepSeek-V3.2 is accessed via the DeepSeek API. For non-thinking inference, we use the model `deepseek-chat` with `temperature=0`. For thinking inference, we use `deepseek-reasoner`. This mode produces an internal chain-of-thought in the `reasoning_content` field before the final answer. In both settings, we request JSON output with `response_format` set to `{"type":"json_object"}`.

Grok-4-Fast is accessed via the xAI batch API. Requests are chunked into batches of up to 50,000 items, uploaded as JSONL, and retrieved after completion. The non-reasoning configuration uses `grok-4-fast`; the reasoning configuration uses `grok-4-fast-reasoning` with extended thinking tokens. All requests use `temperature=0`.

B Additional Results

B.1 Full Results Matrix

Table 7 reports the complete 32-configuration comparison on the official test set.

B.2 Qwen3 Thinking-Mode Diagnostic

To probe whether disabling native Qwen thinking mode made the main PEFT comparison overly pessimistic, we trained a dedicated Qwen3-32B checkpoint under the baseline setup and compared inference with `enable_thinking=False` versus `enable_thinking=True`. Because thinking-mode inference was approximately 4–5$\times$ slower (about 100 minutes for 100 samples in our setup) and the follow-up used different hardware (RTX PRO 6000 Blackwell rather than H100), we evaluated only stratified samples of 100 instances for the thinking runs. This diagnostic should therefore be read as within-checkpoint evidence about output stability and reasoning support rather than as a direct replacement for the main benchmark.

The non-thinking row in Table 8 refers to this diagnostic checkpoint under CoT JSON output format and is therefore distinct from the best Qwen3-32B configuration in the main paper. Its poor score is driven by output-format failure: 37.99% of responses trigger a hard fallback to *Explicit*. Enabling thinking reduces this to 0.00% on the official-test sample and 1.00% on the blind set (Table 6), indicating that native reasoning materially stabilizes output formatting as well as downstream classification.

C Prompt Templates

We use six distinct prompt templates across our experiments. The templates differ along two dimensions: (1) output format (standard text vs. structured CoT JSON) and (2) taxonomy presentation (hierarchical vs. flat). Below, we reproduce the exact text of each template. Placeholders `{interview_question}`, `{question}`, and `{answer}` are filled at inference time.

C.1 Standard Prompts (Fine-Tuned Models)

Zero-shot Standard. This prompt is used for fine-tuned models with standard output format (`Label: <label>`).

```
You are a political discourse analyst. Your task
is to classify the evasion technique used in
Question-Answer pairs.

## Taxonomy (9 Evasion Labels)
Selected the best fitting label from the list
below:

1. **Explicit**: The respondent directly answers
   the specific question asked with the expected
   information.
2. **Implicit**: The answer is provided but
   requires inference (not stated plainly).
3. **General**: The respondent talks about the
   general topic but offers broad, vague
   platitudes instead of specific details.
4. **Dodging**: The respondent completely ignores
   the question content and shifts to an
```

Model	Mode	Prompt strategy	Evasion F1 (Subtask 2)	Clarity F1 (Subtask 1)
Approach A: Fine-tuned Models				
Qwen3-4B	Base	Zero-shot	0.2302	0.4459
Qwen3-4B	Base	Few-shot	0.1400	0.4203
Qwen3-4B	Base	Zero-shot CoT	0.0902	0.2455
Qwen3-4B	Base	Few-shot CoT	0.2455	0.5682
Qwen3-4B	Instruct	Zero-shot	0.3384	0.5307
Qwen3-4B	Instruct	Few-shot	0.2327	0.5338
Qwen3-4B	Instruct	Zero-shot CoT	0.1064	0.2249
Qwen3-4B	Instruct	Few-shot CoT	0.1738	0.4590
Qwen3-14B	Base	Zero-shot	0.3201	0.5671
Qwen3-14B	Base	Few-shot	0.3150	0.5628
Qwen3-14B	Base	Zero-shot CoT	0.3022	0.5073
Qwen3-14B	Base	Few-shot CoT	0.3122	0.5047
Qwen3-32B	Base	Zero-shot	0.3220	0.5879
Qwen3-32B	Base	Few-shot	0.3215	0.5920
Qwen3-32B	Base	Zero-shot CoT	0.3288	0.6126
Qwen3-32B	Base	Few-shot CoT	0.3630	0.5654
Approach B: API Models				
DeepSeek-V3.2	Non-reasoning	Zero-shot CoT (hier.)	0.2646	0.4998
DeepSeek-V3.2	Non-reasoning	Zero-shot CoT (flat)	0.3588	0.5902
DeepSeek-V3.2	Non-reasoning	Few-shot CoT (hier.)	0.4578	0.7605
DeepSeek-V3.2	Non-reasoning	Few-shot CoT (flat)	0.4092	0.6968
DeepSeek-V3.2	Reasoning	Zero-shot CoT (hier.)	0.3085	0.6030
DeepSeek-V3.2	Reasoning	Zero-shot CoT (flat)	0.3795	0.6275
DeepSeek-V3.2	Reasoning	Few-shot CoT (hier.)	0.5115	0.7735
DeepSeek-V3.2	Reasoning	Few-shot CoT (flat)	0.4714	0.7369
Grok-4-Fast	Non-reasoning	Zero-shot CoT (hier.)	0.2636	0.4909
Grok-4-Fast	Non-reasoning	Zero-shot CoT (flat)	0.4075	0.6746
Grok-4-Fast	Non-reasoning	Few-shot CoT (hier.)	0.4071	0.7088
Grok-4-Fast	Non-reasoning	Few-shot CoT (flat)	0.4157	0.6742
Grok-4-Fast	Reasoning	Zero-shot CoT (hier.)	0.3015	0.6470
Grok-4-Fast	Reasoning	Zero-shot CoT (flat)	0.5126	0.7821
Grok-4-Fast	Reasoning	Few-shot CoT (hier.)	**0.5147**	**0.7979**
Grok-4-Fast	Reasoning	Few-shot CoT (flat)	0.5043	0.7565

Table 7: Macro F1 for all 32 model–prompt configurations on Subtask 2 and Subtask 1 on the official test set (N = 308). The overall best result is in **bold**; the best per-group result is underlined.

Split	Mode	N	Evasion F1	Clarity F1
Official Test	Non-thinking	308	0.0734	0.2459
Official Test	Thinking	100	**0.2362**	**0.5127**
Blind Set	Thinking	100	**0.3700**	**0.5913**

Table 8: Diagnostic thinking-mode follow-up for Qwen3-32B. Non-thinking is evaluated on the full official test set; thinking is evaluated on stratified samples due to compute constraints, so direct cross-N comparisons should be interpreted cautiously.

```
   unrelated topic.
5. **Deflection**: The respondent acknowledges
   the topic but shifts focus to a specific
   tangent (e.g., attacking, self-praise) to
   avoid the core inquiry.
6. **Partial/half-answer**: The respondent
   addresses only one part of a multi-part
   question or a sub-issue, ignoring the main
   point.
7. **Declining to answer**: Explicit refusal to
   answer or stating they cannot comment.
8. **Claims ignorance**: The respondent states
   they do not know the answer or lack
   information.
9. **Clarification**: The respondent answers with
   a question to clarify context or meaning.

## Classification Guidelines
1. Check if answer directly addresses what
   question requests.
2. Explicit = direct answer. All others are
   evasion forms.
3. Distinguish: Implicit (implied) vs General
   (vague) vs Dodging (unrelated) vs Deflection
   (shifted focus).

## Output Format
Return exactly:
Label: <label>

## Analysis Task
```

```
Full Interview Context: {interview_question}

Specific Question: {question}

Answer: {answer}

Classify the evasion technique.
```

Few-shot Standard. Extends the zero-shot template by inserting six exemplars (one per representative label) between the taxonomy and the analysis task. Each exemplar contains a question, answer, and gold label. Due to space constraints, we show one representative exemplar below; the full set covers *Explicit*, *Deflection*, *Partial/half-answer*, *General*, *Implicit*, and *Claims ignorance*.

```
### Example 1: Explicit
Question: How would you respond to accusations
  of containing China while pushing for
  diplomatic talks?
Answer: I don't want to contain China. I just
  want to make sure we have a relationship that
  is on the up and up.
Label: Explicit
```

C.2 CoT Prompts with Hierarchical Taxonomy (API Models)

Zero-shot CoT (Hierarchical). Used for API models. Includes a structured taxonomy grouped by clarity tier and requires JSON output with reasoning.

```
You are an expert political discourse analyst
specializing in detecting equivocation and
evasion strategies in high-stakes interviews.
Your objective is to analyze the relationship
between a specific question and the respondent's
answer to classify the evasion technique used.

## The Taxonomy of Evasion
   (Classify into exactly ONE label)

### 1. Clear Reply
* **Explicit**: The respondent directly answers
  the specific question asked with the expected
  information.

### 2. Ambivalent Reply (Evasive or Vague)
* **Implicit**: The answer is provided but
  requires inference.
* **General**: The respondent talks about the
  general topic but offers broad, vague
  platitudes.
* **Partial/half-answer**: The respondent answers
  only one part of a multi-part question.
* **Dodging**: The respondent completely ignores
  the question content.
* **Deflection**: The respondent acknowledges the
  topic but pivots to a specific tangent.

### 3. Clear Non-Reply
* **Declining to answer**: Explicit refusal to
  answer.
* **Claims ignorance**: States they do not know.
* **Clarification**: Asks for clarification.

## Analysis Rules
1. **Directness Check**: Does the answer
   linguistically satisfy the interrogative
   specific of the question?
2. **Topic Fidelity**: Does the answer stay on
   the specific sub-topic of the question, or
   does it drift to a general theme?
3. **Reasoning First**: You must write out your
   analysis *before* deciding the label.

## Output Format
You must return a valid JSON object. Do not
include markdown formatting (like ```json).
Structure:
{
  "id": "string",
  "reasoning": "1. Directness Analysis: ...
    2. Topic Fidelity Analysis: ...
    3. Conclusion: ...",
  "evasion_label": "Exact Label String"
}

## YOUR TASK
Please analyze this specific QA pair:

[Full Context]: {interview_question}
(Note: Use this to understand the broader topic)

[Target Question]: {question}
(Note: Analyze the answer strictly against THIS
specific question)

[Respondent's Answer]: {answer}

Perform the analysis and return the JSON object.
```

Few-shot CoT (Hierarchical). Extends the zero-shot CoT template by inserting nine exemplars (one per label) with full reasoning traces. Each exemplar includes the target question, answer, and a complete JSON output containing directness analysis, topic fidelity analysis, and conclusion. Due to space, we show one exemplar:

```
### Example 5: Dodging (Complete Topic Switch)
[Target Question]: Are you worried about the
  meeting between President Putin and Kim Jong
  Un, if that could mean Russia has more gains
  in the war in Ukraine?
[Respondent's Answer]: Look, I think China has a
  difficult economic problem right now ...
  we're not looking to hurt China, sincerely.
[JSON Output]:
{
  "id": "ex_5_dodging",
  "reasoning": "1. Directness Analysis: The
    question is about Putin, Kim Jong Un, and
    Ukraine. 2. Topic Fidelity Analysis: The
    answer talks entirely about China, Taiwan.
    Zero connection to Russia or North Korea.
    3. Conclusion: Complete topic switch.",
  "evasion_label": "Dodging"
}
```

C.3 CoT Prompts with Flat Taxonomy (API Models)

Zero-shot CoT (Flat / No Taxonomy). Identical structure to the hierarchical variant, but presents labels as a flat numbered list without grouping under clarity tiers. The key difference is the taxonomy section:

```
## Taxonomy (9 Evasion Labels)
Selected the best fitting label from the list
below:

1. **Explicit**: The respondent directly answers
   the specific question asked with the expected
   information.
2. **Implicit**: The answer is provided but
   requires inference (not stated plainly).
3. **General**: The respondent talks about the
   general topic but offers broad, vague
   platitudes instead of specific details.
4. **Dodging**: The respondent completely ignores
   the question content and shifts to an
   unrelated topic.
5. **Deflection**: The respondent acknowledges
   the topic but shifts focus to a specific
   tangent to avoid the core inquiry.
6. **Partial/half-answer**: The respondent
   addresses only one part of a multi-part
   question or a sub-issue, ignoring the main
   point.
7. **Declining to answer**: Explicit refusal to
   answer or stating they cannot comment.
8. **Claims ignorance**: The respondent states
   they do not know the answer or lack
   information.
9. **Clarification**: The respondent answers with
   a question to clarify context or meaning.
```

The analysis rules, output format, and few-shot exemplars remain identical to the hierarchical version. Only the taxonomy presentation differs.

Few-shot CoT (Flat). Combines the flat taxonomy above with the same nine exemplars and reasoning traces as the hierarchical few-shot variant.

D Computational Resources

All fine-tuning experiments were conducted on a single NVIDIA H100 GPU (80 GB VRAM). Using the Unsloth framework, parameter-efficient fine-tuning was exceptionally fast, with the maximum training time for the largest model (32B) taking under 18 minutes. However, the inference process was significantly slower due to the high volume of generation required for the reasoning outputs, with the few-shot chain-of-thought (CoT) prompting configuration being the most time-consuming to evaluate.

API Experiments. API experiments were executed via asynchronous batch endpoints with no local GPU requirements. For Grok-4-Fast, requests were submitted through the xAI Batch API in chunks of up to 25,000 items; each chunk typically completed within 15 minutes, with a worst-case observed latency of under 30 minutes per batch. For DeepSeek-V3.2, requests were submitted via the DeepSeek batch API with similarly low latency.

API Cost. We compute per-sample costs from the `cost_in_usd_ticks` field recorded in all Grok batch result files, calibrated against a known reference sample ($0.0006 for one reasoning request). Table 9 reports the actual cost for all 8 Grok configurations (N = 308 each). Reasoning-mode requests average $0.000411 per sample versus $0.000178 for non-reasoning, a ratio of approximately 2.3×. The best-performing configuration (Grok-4-Fast Reasoning, Few-shot CoT) costs $0.1359 for the full test set. The total expenditure for all 8 Grok configurations is $0.70. DeepSeek-V3.2 API costs are of a similar order of magnitude, giving a total API expenditure across all 16 configurations of approximately $1.50.

Config	Mode	Total ($)	Per sample ($)
Zero-shot flat	Non-reasoning	0.0417	0.000135
Zero-shot hier.	Non-reasoning	0.0461	0.000150
Few-shot flat	Non-reasoning	0.0647	0.000210
Few-shot hier.	Non-reasoning	0.0672	0.000218
Zero-shot flat	Reasoning	0.1148	0.000373
Zero-shot hier.	Reasoning	0.1158	0.000376
Few-shot flat	Reasoning	0.1175	0.000455
Few-shot hier.	Reasoning	0.1359	0.000441
Total		**0.7036**	

Table 9: Actual API cost for all 8 Grok-4-Fast configurations on the official test set (N = 308), derived from `cost_in_usd_ticks` in batch result files.

E Confusion Matrix and Qualitative Error Analysis

Figure 3 presents the full confusion matrix for Subtask 2 using our best configuration (Grok-4-Fast Reasoning, Few-shot CoT) on the official test set (N = 308).

E.1 Case Study Discussion: Reasoning Traces vs. Human Pragmatics

To understand why the best-performing models still struggle with ambivalent evasion categories, we analyze the explicit reasoning traces generated by Grok-4-Fast. Tables 10 and 11 present two case studies where the model's logical deduction cleanly misaligned with human pragmatic judgment.

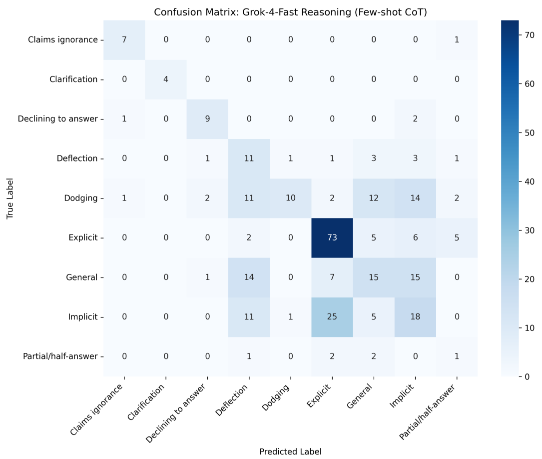

Figure 3: Confusion matrix of Subtask 2 predictions versus gold labels.

Case Study 1: Dodging vs. Deflection

Q: Do you think the Republican leader in the House owes somebody an apology?

A: I wouldn't have exactly put it that way. But I do believe there's a difference of attitude. I mean, take the PATRIOT Act, for example [...] I'm confident the leader, you know, meant nothing personal.

Gold: *Dodging* **Pred:** *Deflection*

Reasoning: *1. Directness:* The question is yes/no about whether the leader owes an apology. The respondent does not explicitly affirm or deny. *2. Topic Fidelity:* Acknowledges controversy but pivots to PATRIOT Act, shifting focus to broader policy differences. *3. Conclusion:* Acknowledges topic but deflects via policy tangent.

Table 10: Model misclassifies complete topic replacement (*Dodging*) as acknowledged pivot (*Deflection*).

E.1.1 Case Study 1: Dodging vs. Deflection Confusion

The first common failure mode involves distinguishing between *Dodging* (complete topic switch) and *Deflection* (acknowledged topic with tangential pivot). Table 10 illustrates this challenge. The model successfully detects the pivot to the PATRIOT Act policy discussion. However, it fails to weight the *severity* of the pivot correctly. Because the politician mentions "difference of attitude" and "meant nothing personal", the model assumes the original topic (the apology demand) was sufficiently acknowledged, thus assigning *Deflection*.

Human annotators, conversely, recognized that the ensuing 200-word policy defense entirely *replaced* the apology demand rather than merely deflecting from it, constituting a complete *Dodge*. This reveals a fundamental limitation: the model's decompositional reasoning breaks down pragmatic judgment into discrete logical steps, losing the holistic assessment of response adequacy that humans perform naturally. The mere mention of the original topic's semantic field (political disagreement) is insufficient to constitute topic acknowledgment if the core question (about a specific apology) is never addressed.

E.1.2 Case Study 2: General vs. Implicit Confusion

The second failure mode involves the boundary between *General* (vague platitudes) and *Implicit* (answer requiring inference). Table 11 demonstrates this challenge. The model logically extracts positive sentiment from the respondent's mention of "priority lists" and "alliances", treating these cues as a solvable puzzle that yields an *Implicit* answer (i.e., "No, I am not concerned because the region is a priority").

However, human annotators interpreted these same statements as political platitudes that offered no concrete answer regarding actual fiscal pressures, labeling it *General*. The critical distinction lies in *specificity*: an implicit answer still addresses the question's core inquiry through contextual cues, whereas a general answer substitutes broad rhetoric for substantive response. The model's step-by-step deduction over-extracts specific intent from diffuse political language, failing to recognize when semantic content is merely decorative rather than informative.

Implications for Future Work: These case studies reveal that chain-of-thought reasoning, while improving overall accuracy, introduces systematic biases toward over-interpretation. The model's explicit logical decomposition paradoxically undermines its ability to recognize absence of substance—a skill that requires holistic pragmatic judgment rather than compositional semantic analysis. Future approaches might benefit from incorporating adversarial training examples that specifically target these boundary cases, or developing hybrid architectures that combine logical decomposition with gestalt-level assessment of response adequacy.

E.1.3 Case Study 3: Model Prediction Aligns with Minority Annotator

The preceding case studies illustrate model failures. However, annotator disagreement also creates cases where a model's prediction, while diverging from

Case Study 2: General vs. Implicit
Q: Are you concerned that fiscal pressures may compromise US plans for the region?
A: With respect to Europe, I'm deeply concerned [...] with respect to our budget, there's a reason why I'm spending this time out here in Asia [...] President Obama being here shows he values the US role in this region and our alliance [...].
Gold: *General* **Pred:** *Implicit*
Reasoning: *1. Directness:* The question asks yes/no about fiscal concerns. Respondent doesn't state concern but provides reassurance by emphasizing regional priority. *2. Topic Fidelity:* Transitions to budget and Asia-Pacific commitment, staying relevant. *3. Conclusion:* Answer implies lack of concern through priority statements—requires inference, fitting *Implicit*.

Table 11: Model over-extracts specific intent from platitudes, misclassifying *General* as *Implicit*.

the gold label, is arguably more defensible than the majority annotation. We present one such instance to contextualize the inherent subjectivity of the task.

Table 12 shows a QA pair where two of three annotators labeled the response *Dodging*, establishing the gold label, while the third annotator labeled it *Implicit*. The model (Grok-4-Fast Reasoning, Few-shot CoT) predicted *Implicit*, aligning with the minority annotator.

The response "I'm going to stay out of Connecticut" does not explicitly answer the yes/no question about campaigning. However, it is *topically anchored*: Connecticut is Senator Lieberman's home state and the location where any such campaign would take place. The geographic reference serves as an indirect indexical that conveys a clear position (non-involvement) through inference—the defining criterion of *Implicit*. By contrast, *Dodging* requires a *complete* topic switch with no connection to the original question; here the connection is direct, even if compressed.

The majority annotation of *Dodging* likely reflects annotators prioritizing the absence of an explicit yes/no commitment and the brevity of the response, rather than the semantic content of the geographic anchor. This case illustrates that some majority-vote labels are not uniquely determined under the taxonomy, and that model errors computed against a single gold label may include pragmatically defensible alternatives.

Case Study 3: Annotator Disagreement — *Implicit* vs. *Dodging*
Q: Would you campaign against Senator Joe Lieberman on Iraq?
A: I'm going to stay out of Connecticut.
Gold: *Dodging* **Pred:** *Implicit* **Annotators:** *Dodging / Dodging / Implicit*
Model Reasoning: *1. Directness:* The question asks yes/no about campaigning. The answer does not explicitly state yes or no. *2. Topic Fidelity:* The response references Connecticut, Lieberman's home state and the site of any such campaign, directly tying to the question context. *3. Conclusion:* Staying out of Connecticut implies non-involvement in campaigns there—an answer conveyed through geographic inference, fitting *Implicit*.

Table 12: A case where the model's *Implicit* prediction aligns with the minority annotator. The geographic anchor (Connecticut = Lieberman's state) constitutes topical relevance, making *Dodging* (complete topic switch) a debatable majority label under the taxonomy definition.