AsymVerify at SemEval-2026 Task 6: Asymmetric Confidence-Gated Verification for Political Evasion Detection

Sebastien Kawada
Kaons K^*
Los Angeles, CA, USA
sebastien@kaons.com

Abstract

Political evasion is difficult to detect because evasive answers often appear cooperative while avoiding concrete commitment. We present AsymVerify, a confidence-gated verification system for SemEval-2026 Task 6, a three-way classification of Clear Reply, Ambivalent, and Clear Non-Reply responses. AsymVerify scored 0.85 Macro F1 on the evaluation split (D_{eval}, n=237), placing **2nd out of 41 teams** on the official leaderboard. The system first classifies each question-answer pair, then selectively applies downgrade verification (CR/CNR $\rightarrow$ AMB) or upgrade verification (AMB $\rightarrow$ CR) to low-confidence predictions. Development analysis shows that errors concentrate at the Ambivalent boundary in both directions, motivating this asymmetric two-verifier design while confidence gating keeps additional inference cost low. On D_{dev} (n=308), AsymVerify with GLM-4.7 gains +17.1 Macro F1 over single-pass classification at 1.48 calls/example, and the upgrade verifier alone improves every tested LLM backend on D_{dev} by +6.8 to +15.2 Macro F1 over its single-pass baseline. Code is available at https://github.com/kaons-research/AsymVerify-ACL.

1 Introduction

Viewers often struggle to detect political question dodges without explicit cues (Clementson, 2018b,a). Large language models may face a related failure mode: preference-optimized systems can favor agreeable or user-aligned interpretations over truthfulness (Sharma et al., 2024), the opposite of what evasion detection requires. Politicians exploit this gap with rhetorical strategies that technically engage questions while avoiding concrete commitment (Bull and Mayer, 1993). Skilled evasion appears substantive, employing topic-adjacent responses that feel informative while committing to nothing.

SemEval-2026 Task 6 requires classifying political question-answer pairs into Clear Reply (CR), Ambivalent (AMB), or Clear Non-Reply (CNR) (Thomas et al., 2026). The task builds on the QEvasion taxonomy and corpus (Thomas et al., 2024), and the CLARITY release provides 3,448 training examples with 59% Ambivalent responses. The dataset consists of English-language U.S. political interview transcripts.

We evaluate on three dataset splits: D_{train} (3,448 examples) for prompt development, D_{dev} (308) for threshold tuning and ablations, and D_{eval} (237) for official placement. On D_{eval}, the GPT-5.2 AsymVerify submission scored 0.85 on CLARITY Subtask 1 and placed **2nd out of 41 teams**. On D_{dev}, model-family analyses examine the mechanism across multiple LLM backends. Confidence analysis shows a clear accuracy gap between high- and low-confidence predictions, which is the basis for selective verification. We further observe that CR and CNR, as opposite speech acts, are rarely confused directly (3% of errors) and are instead typically misclassified as AMB. Because both verification directions correct through AMB, either pass alone targets a large share of errors.

2 Background

2.1 Task Definition

The task defines three response clarity classes:

- **Clear Reply (CR)** — direct answer with specific commitment.
- **Ambivalent (AMB)** — evasive through vague language, topic shifts, or implicit answers.
- **Clear Non-Reply (CNR)** — explicit refusal or claim of ignorance.

Table 1 shows dataset statistics. The dev set has higher Ambivalent proportion (67%) than training (59%), making minority-class Macro F1 harder. Systems are evaluated by Macro F1, the unweighted mean $(\text{F1}_{\text{CR}} + \text{F1}_{\text{AMB}} + \text{F1}_{\text{CNR}})/3$,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1554–1567
July 3-4, 2026

Split	N	AMB	CR	CNR
Train	3,448	59%	31%	10%
Dev	308	67%	26%	8%
Eval	237	*(official placement)*		

Table 1: Dataset splits. Dev set skews toward Ambivalent (67% vs 59%).

a standard choice for imbalanced multi-class classification because each class contributes equally to the final score (Sokolova and Lapalme, 2009). We report accuracy as a secondary metric.

Dataset notation. To avoid cross-split ambiguity, we use D_{train} (n=3,448), D_{dev} (n=308), and D_{eval} (n=237). Unless stated otherwise, prompt tuning and ablations are reported on D_{dev}, while official placement claims use D_{eval}.

2.2 Related Work

Bull and Mayer (1993) typologize tactics for eluding questions in political interviews, while Clayman (2001) shows how interviewees can resist questions while preserving the norm of answering. Bavelas et al. (1990) introduced the Situational Theory of Communicative Conflict, arguing that equivocation arises from communicative dilemmas where all direct responses carry negative consequences. Recent political-QA work measures answer quality through question-answer fit (Alvarez and Morrier, 2026). Our task differs from stance detection (Mohammad et al., 2016), which classifies opinion direction rather than response clarity, because evasion detection requires reasoning about whether commitments were made, not what position was taken.

Unlike self-consistency (Wang et al., 2023), which improves accuracy by voting over multiple LLM samples, AsymVerify uses confidence to route examples selectively into verifier calls. This follows recent work on language-model cascades (Gupta et al., 2024), adaptive computation (Schuster et al., 2022), and cost-aware LLM routing (Chen et al., 2023) that route inputs through different computational paths based on difficulty. Recent text-classification benchmarks also warn that heavier reasoning strategies can impose large token costs without uniform classification gains (Guo et al., 2026), so our verifier design applies extra reasoning only to uncertain boundary cases.

Our verification passes can be viewed as a task-specific form of iterative self-refinement (Madaan et al., 2023; Shinn et al., 2023), where re-examination is asymmetrically conditioned on the initial prediction class rather than using open-ended self-correction, which can be unreliable without external feedback (Huang et al., 2024a). Structured reasoning has shown promise for implicit subjective classification (Sun et al., 2023; Fei et al., 2023), where surface-level features are insufficient. Evasion detection has the same property.

Verbalized confidence tends toward overconfidence (Groot and Valdenegro-Toro, 2024), but recent work shows that prompted confidence can still provide useful uncertainty information when interpreted carefully (Xiong et al., 2024; Huang et al., 2024b). Newer uncertainty-estimation work distinguishes whether an uncertainty signal identifies hard cases from whether it is probabilistically calibrated (Li et al., 2025), and AsymVerify uses confidence in the former sense, as a routing cue. RLHF can further degrade calibration, though verbalized confidence partially recovers it (Tian et al., 2023), and systematic sycophancy in RLHF-trained models causes them to prefer agreeable interpretations over accurate ones (Sharma et al., 2024). This is particularly problematic for evasion detection, where models must resist accepting vague answers as clear. Our prompt design counters this tendency through structured taxonomy guidance and verification prompts that instruct skepticism.

3 System Description

AsymVerify operates in up to three passes, as shown in Figure 1, with pseudocode in Appendix A. The base classifier returns $(\hat{y}, c) = f_\theta(q, r)$ where $\hat{y} \in \{\text{CR}, \text{AMB}, \text{CNR}\}$ and $c \in [0, 1]$ is verbalized confidence, a self-reported scalar from structured JSON output rather than token logprobs (Tian et al., 2023; Xiong et al., 2024). High-confidence predictions ($c \geq \tau$) exit immediately, in the spirit of cascade methods that defer harder instances and resolve easier ones with fewer calls (Gupta et al., 2024). In the reported D_{dev} analyses, low-confidence predictions keep the single Pass 1 label and receive targeted verification rather than maj@3 voting.

The base classifier prompts the selected model with a structured evasion taxonomy containing nine response subtypes (explicit, implicit, general, partial, dodging, deflection, declining, claims ignorance, clarification). The official submission used GPT-5.2, while development analysis uses GLM-

Stage 1: Base Classification

$$(\hat{y}_1, c_1) = f_\theta(q, r), \quad \hat{y}_1 \in \mathcal{Y}$$
$$\mathcal{Y} = \{\text{CR}, \text{AMB}, \text{CNR}\}$$

Stage 2: Confidence Gate (single candidate)

$$\hat{y} = \hat{y}_1 \quad (c_1 \geq \tau;\ \text{exit}),$$
$$\tilde{y} = \hat{y}_1 \quad (c_1 < \tau;\ \text{verify}).$$

Budget shown: single-candidate routing.

Stage 3: Conditional Verification ($c_1 < \tau$)

$$\mathcal{E} = \{\text{CR}, \text{CNR}\},$$
$$y^{\downarrow} = \begin{cases} \text{AMB}, & \tilde{y} \in \mathcal{E},\ g_{\downarrow}(q, r, \tilde{y}) = 1, \\ \tilde{y}, & \text{otherwise}, \end{cases}$$
$$\hat{y} = \begin{cases} \text{CR}, & y^{\downarrow} = \text{AMB},\ g_{\uparrow}(q, r) = 1, \\ y^{\downarrow}, & \text{otherwise}. \end{cases}$$

Routing profile

52.6%	29	120
early exit	P2 calls	P3 calls

1.48 calls/example

Figure 1: AsymVerify control flow. The D_{dev} call budget uses single-candidate routing rather than maj@3. P2 can downgrade CR/CNR predictions to AMB before P3 optionally upgrades AMB predictions to CR. Routing counts are from GLM-4.7 on D_{dev} (n=308; Table 8).

4.7, DeepSeek-V3.2, and Llama-3.3-70B on D_{dev}. The prompt emphasizes concrete commitments and instructs skepticism toward answers that sound substantive but avoid the specific question. The model returns structured JSON with label, confidence $c \in [0, 1]$, and reasoning. All reported D_{dev} call budgets use the single-candidate pipeline: each example receives one base classification call, and low-confidence examples are sent directly to the conditional verifiers. A maj@3 variant can replace the Pass 1 label, but it is not the cost setting reported in Table 8.

Pass 2 re-examines CR and CNR predictions for possible downgrade to AMB using a "one vs. multiple interpretations" criterion, where if reasonable readers could disagree about what was actually said, the response is Ambivalent. Pragmatic accounts (Grice, 1975; Bavelas et al., 1990; Clayman, 2001) treat indirect answers as satisfying conversational relevance without making an explicit commitment. Pass 3 re-examines AMB predictions for possible upgrade to CR by checking whether the first substantive sentence directly answers the question while ignoring preambles and later tangents, a heuristic motivated by the turn-by-turn structure of news interviews (Clayman and Heritage, 2002; Bull and Mayer, 1993). We do not include an AMB→CNR upgrade pass because CNR comprises only 8% of the dataset and CNR→AMB errors account for just 5% of total errors (Table 5), offering minimal recovery potential. Both verification passes run only on low-confidence predictions, and P3 runs on any low-confidence example whose current label is AMB after P2, which explains why verifier calls can slightly exceed the number of low-confidence examples.

Prefilter for upgrade candidates. Pass 3 can be expensive because 67% of D_{dev} examples are AMB. We therefore analyze, but do not use in the main GLM ablation, a rule-based prefilter that selects only AMB predictions whose first sentence shows strong commitment signals before sending them to the LLM verifier. Four lexical rules trigger verification: (1) answers starting with "No" followed by a short declarative sentence, (2) answers starting with "Because" (direct causal explanation), (3) answers starting with "That is" or "That's" (declarative assertion), and (4) the pattern "No, I don't see" (stance-taking). Applied to the GLM-4.7 full-system decisions, this prefilter reduced P3 calls from 120 to 3 and false upgrades from 19 to 0 (Table 10).

The class error distribution further constrains verification design. CR and CNR are opposite speech acts that are rarely confused directly (3% of errors) and are instead typically misclassified as AMB, so all corrections route through AMB.

4 Experimental Setup

Data. We use D_{train} (3,448) for prompt development, D_{dev} (308) for threshold tuning and ablations, and D_{eval} (237) for official placement. During submission, D_{eval} labels were hidden, so model selection and analysis used D_{dev}.

Models. The official submission used GPT-5.2 with `reasoning_effort=high` (OpenAI, 2025). Development-set ablations use GLM-4.7 as the primary analysis model (OpenRouter, 2026), and model-family replication evaluates DeepSeek-V3.2 (DeepSeek-AI, 2025) and Llama-3.3-70B (Meta AI, 2024) on D_{dev}. We report GPT-5.2 for the official evaluation split and use the development models for pass-level analysis, reflecting recent findings that model family, prompting strat-

egy, and inference cost materially affect LLM text-classification behavior (Kostina et al., 2025; Guo et al., 2026).

Hyperparameters. Confidence threshold $\tau = 0.95$ selected from $\{0.85, 0.90, 0.95, 1.0\}$ on D_{dev}. Temperature is 0.1 for base classification and 0.0 for verification. Because LLM confidence measures can be miscalibrated even when they are useful for ranking difficulty (Groot and Valdenegro-Toro, 2024; Huang et al., 2024b; Tian et al., 2023; Li et al., 2025), we treat τ as a routing threshold rather than a calibrated probability. Each configuration uses one low-temperature decoding pass, and we add 1,000-sample paired bootstrap intervals for the close deltas most likely to be sampling-sensitive (Koehn, 2004; Dror et al., 2018).

5 Results

We first report official leaderboard performance on D_{eval} for CLARITY Subtask 1 from the shared-task overview (Thomas et al., 2026), then analyze the mechanism on D_{dev}. Table 2 gives the top-12 leaderboard.

Rank	Participant	Prediction score
1	TeleAI	**0.89**
2	**AsymVerify**	**0.85**
3	CSE-UOI	0.85
4	Rasende Rakete	0.83
5	Evaluators	0.83
6	YNU-HPCC	0.83
7	moswisarut	0.82
8	tahamunawar	0.81
9	CLaC @ CLARITY	0.80
10	SpinDetector	0.80
11	gabriel_stefan	0.80
12	AGAI	0.79

Table 2: Official CLARITY Subtask 1 leaderboard on D_{eval} (n=237; 41 teams). Scores are official prediction scores, and AsymVerify placed **2nd out of 41 teams**, with CSE-UOI also displaying 0.85.

5.1 Development Analysis (Dev Set)

Table 3 tests whether the verification mechanism is tied to a single backend. The P3 upgrade branch improves Macro F1 by +6.8 to +15.2 across all three D_{dev} backends (GLM-4.7, DeepSeek-V3.2, Llama-3.3-70B). P2 is more model-sensitive: it gives the best GLM-4.7 result but over-corrects DeepSeek and Llama, indicating that downgrade verification needs model-specific thresholding.

Model	Config.	F1	Acc.	Calls/ex.
GLM-4.7	P1	55.9%	77.6%	1.00
GLM-4.7	P1+P3	70.8%	72.1%	1.33
GLM-4.7	P1+P2+P3	**73.0%**	75.3%	1.48
DeepSeek-V3.2	P1	55.9%	73.4%	1.00
DeepSeek-V3.2	P1+P3	**62.7%**	72.1%	1.66
DeepSeek-V3.2	P1+P2+P3	55.3%	67.2%	1.70
Llama-3.3-70B	P1	41.0%	70.8%	1.00
Llama-3.3-70B	P1+P3	**56.3%**	74.4%	1.94
Llama-3.3-70B	P1+P2+P3	53.7%	73.1%	1.97

Table 3: Model-family replication on D_{dev} (n=308). P3-only improves every tested backend, while full P2+P3 is strongest on GLM-4.7 but not uniformly best.

Gold	Pred CR	Pred AMB	Pred CNR
CR	56	22	1
AMB	42	158	6
CNR	1	4	18

Table 4: Confusion matrix on D_{dev} (n=308), GLM-4.7 P1+P2+P3.

5.2 Cross-Model Class Stability

Improvements are not uniform across classes. Ambivalent detection remains stable across the best replicated variants (79.8–83.1 F1), suggesting that the evasion-focused prompt template transfers reliably even when the base model changes. In contrast, Clear Reply varies by 11.8 points (51.2–63.0), and Clear Non-Reply varies by 40.5 points (34.5–75.0), indicating that explicit refusal detection is substantially more model-dependent.

This asymmetry explains why portability gains can be large in Macro F1 while overall accuracy remains in a narrow band (72.1–75.3%). Models can converge on majority-class AMB behavior yet still diverge on minority classes, especially CNR. Practically, backend replacement appears less sensitive for AMB-heavy analyses, while CNR-sensitive use cases require model-specific prompt adaptation and threshold re-tuning.

The full confusion matrix in Table 4 provides the bridge from class-level stability to error mechanism: the largest off-diagonal cells are AMB→CR and CR→AMB, while direct CR↔CNR swaps are sparse. The aggregated error distribution in Table 5 makes the pattern explicit. The two dominant errors are AMB→CR (55%), where evasive responses are over-credited as commitments, and CR→AMB (29%), where clear commitments are penalized for rhetorical hedging. Direct CR↔CNR confusion is rare (3%), showing that errors concentrate at the boundaries with AMB rather than between the polar classes.

Error Type	Count	%
AMB → CR	42	55%
CR → AMB	22	29%
AMB → CNR	6	8%
CNR → AMB	4	5%
CR ↔ CNR	2	3%

Table 5: Error patterns on D_{dev} (n=308), GLM-4.7 P1+P2+P3. Direct CR↔CNR confusion is rare because errors concentrate at class boundaries.

Error Family	Count	Share
P2-aligned (AMB→CR/CNR)	48	63.2%
P3-aligned (CR→AMB)	22	28.9%
Outside current passes	6	7.9%

Table 6: Remaining-error alignment by verification direction on D_{dev} (n=308), derived from Table 5 (76 total errors).

5.3 Error Coverage by Verification Direction

This boundary concentration has a practical consequence. Over 90% of remaining errors fall along the two verification directions ($E_{\downarrow}$ for AMB→{CR,CNR} and $E_{\uparrow}$ for CR→AMB, with arrows denoting gold→predicted labels). Table 6 reorganizes the same 76 errors by *which pass direction they align with*. Pass 2 (downgrade) aligns with over-acceptance errors where AMB is predicted as CR/CNR (48 errors, 63.2%). Pass 3 (upgrade) aligns with under-recognized commitments where CR is predicted as AMB (22 errors, 28.9%). Only 6 errors (7.9%) fall outside both routes.

This decomposition explains why single-pass variants already recover most of the available gain. Both passes target large but complementary slices of boundary-concentrated errors, so either direction alone improves strongly (Table 7), while combining both gives a smaller but consistent additional gain.

The ablation in Table 7 quantifies each pass's contribution on GLM-4.7. The gap between base accuracy (77.6%) and Macro F1 (55.9%) reflects the 67% AMB class imbalance, where high accuracy is achievable by defaulting to the majority class while Macro F1 demands balanced performance. Each verification pass independently contributes about +15 points, and combining both yields +17.1. Selective verification achieves this at roughly half the call budget of running all three passes unconditionally (457 versus 924 calls).

Config.	F1	Acc.	Calls/ex.	P2	P3
P1	55.9%	77.6%	1.00	0	0
P1+P2	70.9%	76.0%	1.15	45	0
P1+P3	70.8%	72.1%	1.33	0	102
P1+P2+P3	**73.0%**	75.3%	1.48	29	120

Table 7: Verification pass ablation on D_{dev} (n=308), GLM-4.7. P2/P3 columns report activation counts.

Where do the computational savings come from? In the single-candidate GLM-4.7 full system, confidence gating routes 162 of 308 examples (52.6%) directly to output after Pass 1, leaving 146 low-confidence examples. These examples trigger 149 verifier calls because three P2 downgrades to AMB are subsequently checked by P3. The full system uses $C_{\text{Asym}} = N + n_{P2} + n_{P3} = 457$ total calls versus $C_{\text{all}} = 3N = 924$ for running all three passes unconditionally, a 50.5% reduction. A low-confidence maj@3 variant would add two base calls per low-confidence example, $C_{\text{maj@3}} = C_{\text{Asym}} + 2n_{\text{low}} = 749$.

Stage	Calls	% of full
Pass 1 (all examples)	308	33.3%
Pass 2 + Pass 3 (low-conf only)	149	16.1%
AsymVerify total	**457**	**49.5%**
All 3 passes unconditionally	924	100%

Table 8: API call budget for single-candidate routing on D_{dev} (GLM-4.7, n=308, without maj@3 voting). Confidence gating reduces total calls by 50.5% versus running all passes on every example. A low-confidence maj@3 variant would add two base calls for each of the 146 low-confidence examples, increasing the total from 457 to 749 calls, or from 1.48 to 2.43 calls/example.

The routing threshold is supported by a simple confidence-bin check. In the GLM-4.7 full system, predictions at or above τ=0.95 account for 162 examples, exactly matching the early-exit count, and are more accurate than the aggregate below-threshold set: 80.9% versus 76.0% (Appendix Table 9). The tiny 0.70–0.80 bin is not meaningful on its own, and these bins are not a full rank- or probability-calibration study (Huang et al., 2024b; Li et al., 2025). Instead, the aggregate trend supports verbalized confidence as a routing signal for identifying harder examples.

Appendix A.1 also reports a Pass 3 prefilter analysis. The lexical filter reduces P3 calls from 120 to 3 and eliminates false upgrades (19 to 0), raising Macro F1 from 73.0 to 74.7, and because the rules are English lexical cues, we present this as an efficiency variant rather than the main system.

Paired bootstrap intervals show that close ablation differences should not be over-read, following standard NLP practice for comparing systems on the same examples (Koehn, 2004; Dror et al., 2018). P1+P2 and P1+P3 differ by only -0.1 Macro F1 points, with a 95% CI of [-5.9, 5.9]. The prefilter delta is +1.7 points but its CI includes zero [-0.9, 4.1]. In contrast, Llama's P1→P1+P3 gain is +15.2 points with CI [7.5, 23.1], a robust replication of the upgrade branch.

Appendix E illustrates verification in action with representative success and failure cases from D_{dev}.

6 Discussion

Confidence routing works because high-confidence predictions are usually correct, while low-confidence predictions benefit from targeted verification. Additional calls on easy cases can add noise rather than signal, so confidence acts as a practical difficulty signal that concentrates computation where it matters (Gupta et al., 2024; Schuster et al., 2022).

Different verification strategies converge because errors concentrate at class boundaries rather than spanning opposite classes (Section 5.2). Both downgrade and upgrade verification correct through AMB, and since over-acceptance and under-recognition each account for many errors, either path alone addresses a substantial fraction. The rare CR↔CNR confusions (3%) fall outside both paths but are too infrequent to dominate system behavior.

Residual errors cluster in hedged commitments, procedural refusals, and conditional commitments. These cases are difficult even for humans: QEvasion has Fleiss $\kappa = 0.644$ inter-annotator agreement (Thomas et al., 2024), and political responses are often crafted to appear both decisive and non-committal (Clayman, 2001; Clayman and Heritage, 2002; Clementson, 2018a). Appendix D further shows that semantic embeddings alone provide negligible class separation (silhouette = 0.001), so retrieval and similarity methods are unlikely to help, making prompt-based pragmatic reasoning the relevant alternative.

On D_{dev}, AsymVerify reaches 73.0 Macro F1 using 49.5% of the unconditional three-pass call budget. The control flow is backend-agnostic, with verifier prompts tuned per model. Majority-vote variants can be budgeted separately if used. Future gains will likely require richer discourse-level features or calibrated multi-model disagreement signals rather than additional majority voting alone.

7 Conclusion

AsymVerify scored 0.85 on CLARITY Subtask 1 (D_{eval}), placing **2nd out of 41 teams** on the official leaderboard. Its central result is that verification passes targeting opposite failure modes converge because both flow corrections through the middle Ambivalent class. Confidence gating exploits this convergence to reach 73.0 Macro F1 on GLM-4.7 D_{dev} with 1.48 calls/example. The P3 upgrade branch alone replicates gains of +6.8 to +15.2 Macro F1 across GLM-4.7, DeepSeek-V3.2, and Llama-3.3-70B.

Limitations

Our confidence threshold and verification prompts were optimized on English-language U.S. political interviews, so other political cultures, languages, or interview formats may require re-tuning. Political evasion exists on a spectrum, and QEvasion's $\kappa = 0.644$ agreement means the system inherits judgment calls where "correct" classification can reflect annotator interpretation rather than ground truth. We also do not evaluate adversarially crafted evasions or noisy ASR transcripts.

The official score establishes performance of the GPT-5.2 submission, while pass-level analysis is reported for GLM-4.7, DeepSeek-V3.2, and Llama-3.3-70B on D_{dev}. Results use single low-temperature decoding per configuration, so close ablation differences should be read through the paired bootstrap intervals. Finally, each question-answer pair is classified independently, although multi-turn context, follow-ups, and speaker history may resolve cases that are ambiguous in isolation.

Ethics Statement

AsymVerify is intended to support analysis of public political interviews by identifying whether a response gives a clear answer to the question asked. The labels describe response clarity rather than speaker intent or factual truth, which is important because equivocation can be strategically ambiguous rather than simply false (Bavelas et al., 1990; Clayman, 2001; Clementson, 2018b). System outputs should therefore be interpreted with conversational context and human judgment.

References

R. Michael Alvarez and Jacob Morrier. 2026. Measuring the quality of answers in political Q&As with large language models. *Political Analysis*, 34(1):78–95.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. *Equivocal Communication*. Sage Series in Interpersonal Communication. Sage Publications, Newbury Park, CA.

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

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

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

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

David E Clementson. 2018a. Deceptively dodging questions: A theoretical note on issues of perception and detection. *Discourse & Communication*, 12(5):478–496.

David E Clementson. 2018b. Effects of dodging questions: How politicians escape deception detection and how they get caught. *Journal of Language and Social Psychology*, 37(1):93–113.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Google AI. 2026. Embeddings: Gemini API. Accessed: 2026-05-01.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.

Xinyu Guo, Yazhou Zhang, and Jing Qin. 2026. TextReasoningBench: Does reasoning really improve text classification in large language models? *arXiv preprint arXiv:2603.19558*.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. In *International Conference on Learning Representations*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024b. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*.

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. 2025. Towards harmonized uncertainty estimation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22938–22953, Vienna, Austria. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Meta AI. 2024. Llama 3.3 70B Instruct model card. Hugging Face model card.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

OpenAI. 2025. GPT-5.2 model. OpenAI API documentation. Accessed: 2026-05-03.

OpenRouter. 2026. GLM 4.7: Api pricing and providers. Accessed: 2026-05-01.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pages 17456–17472.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. SemEval-2026 Task 6: CLARITY – unmasking political question evasions. *Preprint*, arXiv:2603.14027.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

A Full Pseudocode

Algorithm 1: ASYMVERIFY. Parameters: τ=0.95. The D_{dev} call budget uses single-candidate routing, with maj@3 treated as a separate variant. P2 and P3 are conditional branches rather than unconditional sequential calls, and $g_{\downarrow}, g_{\uparrow} \in \{0, 1\}$ are LLM verifiers.

Input : Question q, response r
Output : Label $\hat{y} \in \{\text{CR}, \text{AMB}, \text{CNR}\}$
Pass 1: Base classification;
$(\hat{y}, c) \leftarrow f_\theta(q, r)$;
Confidence gating;
if $c \geq \tau$ **then**
  return $\hat{y}$;
Low confidence: single-candidate verifier routing;
$\tilde{y} \leftarrow \hat{y}$;
Conditional branch P2: Downgrade CR/CNR → AMB;
if $\tilde{y} \in \{\text{CR}, \text{CNR}\}$ **then**
  if $g_{\downarrow}(q, r, \tilde{y}) = 1$ **then**
    $\tilde{y} \leftarrow \text{AMB}$;
Conditional branch P3: Upgrade AMB → CR;
if $\tilde{y} = \text{AMB}$ **then**
  if $g_{\uparrow}(q, r) = 1$ **then**
    $\tilde{y} \leftarrow \text{CR}$;
$\hat{y} \leftarrow \tilde{y}$;
return $\hat{y}$;

A.1 Additional Routing Analysis

Tables 9 and 10 report the confidence-bin and prefilter analyses that support the routing decisions summarized in the main results.

Confidence	n	Acc.	Err.
<0.70	16	62.5%	37.5%
0.70–0.80	2	50.0%	50.0%
0.80–0.90	34	76.5%	23.5%
0.90–0.95	94	78.7%	21.3%
≥0.95	162	80.9%	19.1%

Table 9: Verbalized-confidence bin accuracy on D_{dev} (n=308), GLM-4.7 full system.

Variant	F1	Acc.	P3 calls	False up.
No prefilter	73.0%	75.3%	120	19
Lexical prefilter	74.7%	78.9%	3	0

Table 10: Pass 3 prefilter analysis on D_{dev} (n=308), GLM-4.7 full system.

B Prompt Templates

This appendix reproduces the full prompts used in all three passes.

B.1 Base Classification Prompt

Pass 1: Base Classification

Role: You are classifying political Q&A exchanges for evasion.

Guidance

- Politicians are skilled at appearing to answer while actually evading.
- Be skeptical of surface-level cooperation; look for concrete commitments.

Evasion taxonomy *Note:* Answers begin with speaker identification (e.g., "President Trump."). This is transcript formatting; ignore it.

Clear Reply (Explicit): Direct answers providing specific information, a clear yes/no with commitment, or concrete

numbers, names, dates, and policies.

Ambivalent: Evasive responses:

1. Implicit: hints without stating explicitly.
2. General: too vague, lacks specificity.
3. Partial: addresses only part of the question.
4. Dodging: ignores the question or changes topic.
5. Deflection: starts on topic but pivots away.

Clear Non-Reply: Explicit refusal:

1. Declining: explicitly refuses ("I won't comment").
2. Claims ignorance: says they do not know.
3. Clarification: asks for clarification instead.

Task Analyze the exchange: **Question:** "{question}" **Answer:** "{answer}"

Check:

1. What specific information is the question asking for?
2. Does the answer provide that specific information?
3. Is there evasion, deflection, or vagueness?

Output JSON: `{"classification": "Clear Reply" | "Ambivalent" | "Clear Non-Reply", "confidence": 0.0-1.0, "reasoning": "brief"}`

B.2 Downgrade Verification Prompts (Pass 2)

Pass 2a: Clear Reply → Ambivalent

Input Question: "{question}" Answer: "{answer}" *(Skip speaker ID; focus on substantive response.)*

Decision: Does this answer admit only one interpretation or multiple?

Clear Reply: only one interpretation is possible; the answer explicitly commits to a position, no inference needed.
Ambivalent: multiple interpretations are possible; inference is required.

Examples
Q: "Have you seen my chocolates?" A: "The children were in your room this morning."
→ **Ambivalent** (implies the children took them, but does not explicitly say so)
Q: "Have you seen my chocolates?" A: "Yes, they are in the kitchen."
→ **Clear Reply** (only one interpretation)

Output: `{"classification": "Clear Reply" | "Ambivalent", "reasoning": "brief"}`

Pass 2b: Clear Non-Reply → Ambivalent

Input Question: "{question}" Answer: "{answer}" *(Skip speaker ID; focus on substantive response.)*

Decision: Is this a Clear Non-Reply or Ambivalent?

Clear Non-Reply: openly refuses to share information. The refusal is explicit and unambiguous.

- "I don't know" / "I'm not aware" (claims ignorance)
- "I won't comment" / "No comment" (declines)
- "What do you mean?" (asks for clarification)

Ambivalent: provides a response but allows multiple interpretations.

- Leverages the subject to pivot elsewhere (deflection)
- Gives information that does not answer the question
- Appears to engage but does not commit

Examples
Q: "Have you seen my chocolates?" A: "You should not keep chocolates all around the house."
→ **Ambivalent** (deflects; no information about seeing chocolates)
Q: "Have you seen my chocolates?" A: "I don't know where they are."
→ **Clear Non-Reply** (explicit claim of ignorance)

Output: `{"classification": "Clear Non-Reply" | "Ambivalent", "reasoning": "brief"}`

B.3 Upgrade Verification Prompt (Pass 3)

Pass 3: Upgrade Verification

Input Question: "{question}" Answer: "{answer}"
Currently classified as **Ambivalent**. Check if it should be **Clear Reply**. *(Skip speaker ID; inspect first substantive sentence.)*

Upgrade to Clear Reply if the first substantive sentence:

1. Directly answers with yes/no, a specific stance, or a clear position.
2. Does not start with preambles ("Well…", "Look…", "Let me…").
3. Is not immediately followed by "but", "however", or "although".

Important: what comes *after* the first substantive sentence does not matter. The key test: can you extract one clear answer from the opening?

Clear Reply examples: "No, I don't see a contradiction…" (clear stance) "That is one of the options…" (specific commitment) "Because it takes time…" (direct causal)

Stays Ambivalent: "Well, I think…" (preamble) "It depends on…" (conditional) "I wouldn't say…" (negation without stance)

Output: `{"classification": "Clear Reply" | "Ambivalent", "reasoning": "brief"}`

C Extended Model-Family Comparison

Table 11 shows the complete model-family development results.

Model	Config.	Acc.	Macro F1	Calls/ex.
GLM-4.7	P1	77.6%	55.9%	1.00
GLM-4.7	P1+P2	76.0%	70.9%	1.15
GLM-4.7	P1+P3	72.1%	70.8%	1.33
GLM-4.7	P1+P2+P3	75.3%	73.0%	1.48
DeepSeek-V3.2	P1	73.4%	55.9%	1.00
DeepSeek-V3.2	P1+P3	72.1%	62.7%	1.66
DeepSeek-V3.2	P1+P2+P3	67.2%	55.3%	1.70
Llama-3.3-70B	P1	70.8%	41.0%	1.00
Llama-3.3-70B	P1+P3	74.4%	56.3%	1.94
Llama-3.3-70B	P1+P2+P3	73.1%	53.7%	1.97

Table 11: Extended model-family comparison on D_{dev} (n=308).

Model	Config.	CR F1	AMB F1	CNR F1
GLM-4.7	P1+P2+P3	62.9%	81.0%	75.0%
DeepSeek-V3.2	P1+P3	57.0%	79.8%	51.4%
Llama-3.3-70B	P1+P3	51.2%	83.1%	34.5%

Table 12: Per-class F1 for the strongest development variant of each tested backend on D_{dev} (n=308).

Comparison	Δ F1	95% CI
P1+P2 $\rightarrow$ P1+P3	-0.1%	[-5.9%, 5.9%]
P3 no filter $\rightarrow$ prefilter	1.7%	[-0.9%, 4.1%]
DeepSeek P1 $\rightarrow$ P1+P3	6.8%	[-0.1%, 13.2%]
Llama P1 $\rightarrow$ P1+P3	15.2%	[7.5%, 23.1%]

Table 13: Paired bootstrap confidence intervals on D_{dev} (n=308; 1,000 resamples).

D Embedding Space Analysis

To probe whether semantic similarity separates the response-clarity labels, we project 3,756 train+dev embeddings from Gemini Embedding 001 (3,072 dimensions; Google AI, 2026) into two dimensions using UMAP (McInnes et al., 2018). Figure 2 shows near-total class overlap, with a silhouette score of 0.001 indicating essentially no separation in this projection. All three classes intermix throughout the projection, and class centroids are separated by only 0.001–0.016 cosine distance in the projected space. This projection suggests that these embeddings do not cleanly separate response-clarity labels, matching contrastive RAG ablations that did not improve classification. Evasion detection depends on pragmatic cues such as commitment strength, hedging patterns, and rhetorical structure that general-purpose text embeddings did not capture in this setting.

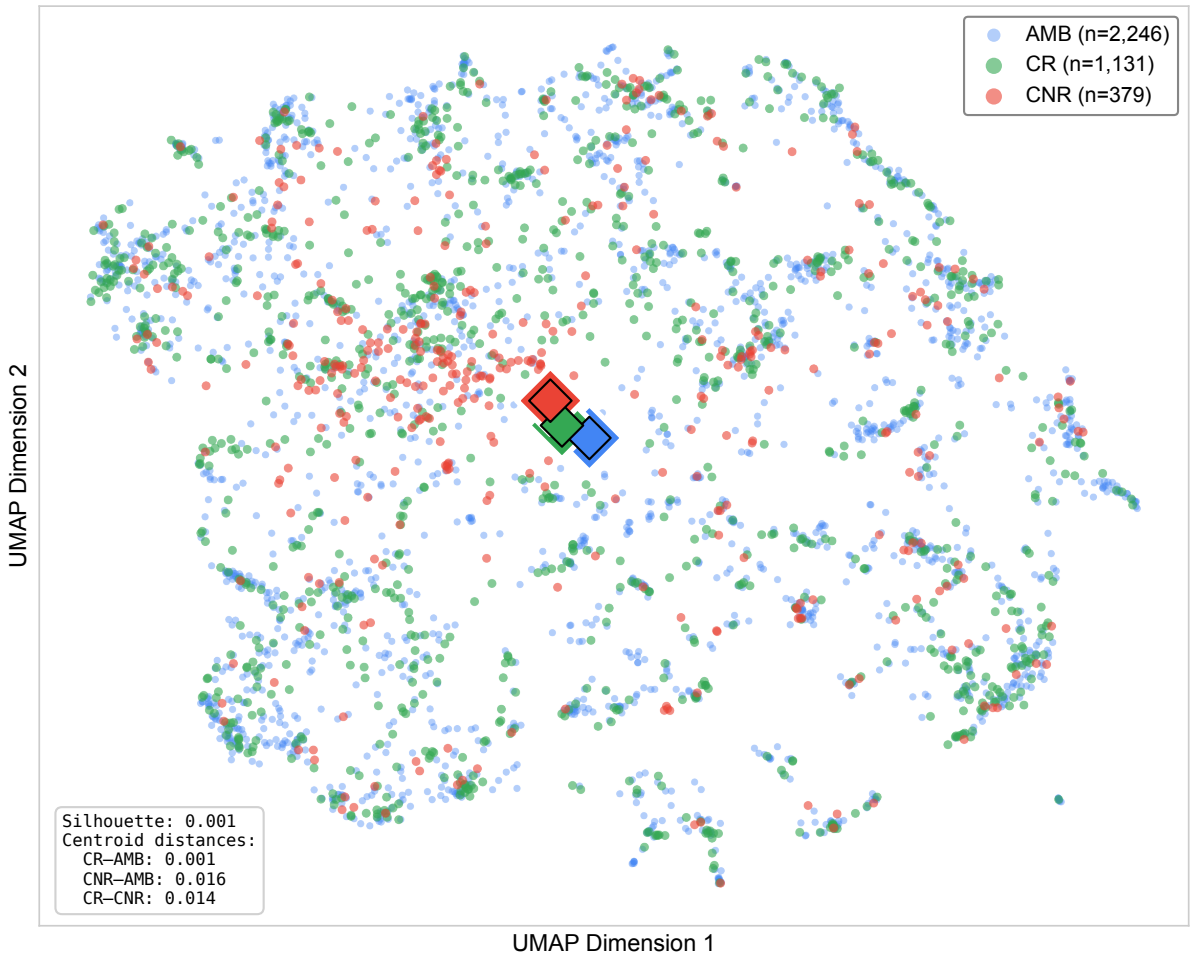

Figure 2: UMAP projection of 3,756 train+dev embeddings (Gemini Embedding 001, 3,072 dimensions). Classes overlap almost entirely (silhouette = 0.001), suggesting that this embedding/projection does not cleanly separate response-clarity labels.

E Extended Qualitative Examples

Tables 14 and 15 show representative predictions with full reasoning traces. We select examples where verification *changed* the initial prediction, demonstrating the pipeline's corrective behavior.

Table 14: Verification saves: cases where Pass 2/3 corrected an initial error.

✓ Pass 3 upgrade: AMB → CR *Gold: CR | Final: CR*

Q: Are you committed to building the 700 miles of fence, actual fencing?
A: "**Yes**, we're going to do both, Joe. We're just going to make sure that we build it in a spot where it works…"

Pass 1 (confidence 0.85 → AMB): "The speaker begins with a direct 'Yes,' which seems like a clear reply, but immediately qualifies the commitment by focusing on building fencing 'in a spot where it works' rather than confirming the specific 700-mile target. The response employs deflection, shifting the focus from the specific mileage to a broader discussion about 'smart borders,' guest worker programs, and terrain conditions."

Pass 3 (upgrade → CR): "The first substantive sentence begins with 'Yes,' directly answering the question, and does not use prohibited preambles or immediate negations like 'but' or 'however'."

Pass 1 was distracted by post-answer qualifications; Pass 3 correctly anchored on the opening "Yes."

✓ Pass 2a downgrade: CR → AMB *Gold: AMB | Final: AMB*

Q: Do you feel like you have been an effective leader when it comes to the economy?
A: "Well, first of all, wherever Republicans indicate an interest…When I came into office, the U.S. economy had contracted by 9 percent…Little over a year later, the economy was growing by 4 percent…**It's not good enough**."

Pass 1 (confidence 0.9 → CR): "The speaker eventually provides a direct, substantive answer by citing specific economic metrics: 'the U.S. economy had contracted by 9 percent…a year later, the economy was growing by 4 percent.' This provides concrete numbers and dates to evaluate effectiveness, fitting the criteria for a Clear Reply."

Pass 2a (downgrade → AMB): "The speaker presents evidence of economic growth but immediately qualifies it by saying it's 'not good enough.' They never explicitly state whether they view themselves as an effective leader, leaving room for multiple interpretations about their self-assessment."

Pass 3 (upgrade check → AMB, no change): "The first substantive sentence starts with the preamble 'Well, first of all' and pivots to discussing Republicans rather than directly answering the specific yes/no question about being an effective leader."

All three passes ran. Pass 1 was swayed by concrete metrics; Pass 2a caught the unanswered yes/no; Pass 3 confirmed the preamble signals evasion.

✓ Pass 3 upgrade after parse failure: AMB → CR *Gold: CR | Final: CR*

Q: Are you still pursuing him?
A: "**Absolutely. No ands, ifs, or buts.** And in my judgment, it's just a matter of time, unless we stop looking…"

Pass 1: Parse failure (no structured output returned). Defaults to AMB with confidence 0.0.

Pass 3 (upgrade → CR): "The first substantive sentence is 'Absolutely,' which directly answers the question with a clear affirmative without preambles or immediate contrasting conjunctions."

The pipeline's fallback-to-AMB + mandatory verification recovered from a complete Pass 1 failure.

✓ Pass 3 upgrade: AMB → CR *Gold: CR | Final: CR*

Q: Is the Vladimir Putin who said "The United States is undermining global security" the same Vladimir Putin whose soul you looked into and found to be trustworthy?
A: "I think **the person who I was referring to in 2001 is the same strong-willed person**. He is a person with whom I have had agreements and disagreements throughout the course of my Presidency…"

Pass 1 (confidence 0.9 → AMB): "The question asks if Putin is still the 'trustworthy' person the speaker referenced in 2001. The speaker acknowledges Putin is the same 'strong-willed' person but deflects to discussing policy disagreements on NATO and cooperation on non-proliferation. He avoids explicitly confirming or denying whether he still finds Putin 'trustworthy.' "

Pass 3 (upgrade → CR): "The first substantive sentence directly answers whether it's the same Putin by stating 'I think the person who I was referring to in 2001 is the same strong-willed person.' This provides a clear position without contradictory qualifiers."

Pass 1 over-indexed on the "trustworthy" framing; Pass 3 recognized that identifying the same person *is* the direct answer.

Table 15: Verification failures: cases where Pass 2/3 introduced or failed to correct errors.

✗ Pass 3 over-upgrade: AMB → CR *Gold: AMB | Final: CR*

Q: Are you still willing to follow a path that seems to be in opposition to the will of the American people?
A: "**I am willing to follow a path that leads to victory**, and that's exactly why we're conducting the review... Victory in Iraq is achievable. It hasn't happened nearly as quickly as I hoped..."

Pass 1 (confidence 0.85 → AMB): "The answer uses deflection and reframing rather than providing a direct 'yes' or 'no' commitment to the question's premise. Instead of acknowledging or explicitly refusing the premise that his path is 'in opposition to the will of the American people,' the speaker redefines the path as one that leads to 'victory' and argues that the public actually supports staying."

Pass 3 (upgrade → CR): "The first substantive sentence ('I am willing to follow a path that leads to victory') directly answers the question with a specific stance, contains no preambles, and is not immediately followed by a negating conjunction."

Failure: Pass 1 correctly identified the reframing ("a path that leads to victory" ≠ the path asked about), but Pass 3's first-sentence heuristic overrode this nuanced judgment.

✗ Pass 2a over-downgrade: CR → AMB *Gold: CR | Final: AMB*

Q: What have they achieved and what will they leave behind for their successors?
A: "Okay, I will start answering. Has it become better, or has it become worse?... Let us remember the world on the brink of a nuclear disaster during the Caribbean crisis... **We have created a reliable platform for cooperation**..."

Pass 1 (confidence 0.9 → CR): "The speaker directly addresses the question by listing specific achievements in the bilateral relationship, such as shifting from enemies to partners and making progress in counterterrorism, non-proliferation, and fighting the drug threat."

Pass 2a (downgrade → AMB): "The answer is vague about specific achievements and doesn't clearly address what will be left behind for successors. While it mentions improved relations, it requires interpretation to understand the concrete nature and significance of these achievements."

Pass 3 (upgrade check → AMB, no change): "The answer begins with the preamble 'Okay, I will start answering,' followed by a rhetorical question ('Has it become better, or has it become worse?'). It fails to provide a direct answer in the first substantive sentence."

Failure: All three passes ran. Pass 1 correctly recognized substantive policy content; Pass 2a and 3 both penalized the rhetorical register despite the answer containing concrete achievements.

✗ Pass 3 over-upgrade: CNR → CR *Gold: CNR | Final: CR*

Q: Why wouldn't you be frustrated, sir, about what's happening?
A: "**I'm not**—I do remember the meeting; I don't remember being surprised. I'm not sure what they meant by that."

Pass 1 (confidence 0.9 → AMB): "The question asks for a reason why the speaker is not frustrated. The answer provides information about memory of a meeting and lack of surprise, but fails to address the specific emotional state of frustration or the reasoning behind it."

Pass 3 (upgrade → CR): "The first substantive sentence ('I'm not—I do remember the meeting; I don't remember being surprised') directly denies the frustration implied in the question and provides a specific factual stance on the meeting without preambles or negating conjunctions."

Failure: The truncated "I'm not—" is a false start, not a denial. The speaker never engages with the frustration question, instead pivoting to meeting recall. This is a refusal by redirection, which neither pass detected.

✗ Pass 3 over-upgrade: AMB → CR *Gold: AMB | Final: CR*

Q: Has the scandal damaged Hastert's credibility and effectiveness in maintaining party control in the midterm elections?
A: "**No**, I think the Speaker's strong statements have made it clear to not only the party members but to the country that he wants to find out the facts..."

Pass 1 (confidence 0.9 → AMB): "The speaker begins with a direct 'No,' but fails to substantively address the specific impact on 'maintaining party control in the midterm elections.' Instead, the response deflects to a defense of the Speaker's statements, general condemnation of the behavior, and a counter-accusation."

Pass 3 (upgrade → CR): "The first substantive sentence begins with 'No,' which is a direct answer to the question, and it is not followed by immediate qualifiers or negations like 'but' or 'however'."

Failure: Pass 1 correctly noted the "No" is followed by topic deflection rather than substantive engagement with the midterm elections premise. Pass 3's first-sentence rule cannot distinguish a genuine "No" from a dismissive one.