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

Sebastien Kawada
Pellucid Research
Los Angeles, CA, USA
sebastien@pellucidresearch.org

Abstract

Political evasion, answering a question without committing to a position, is difficult to detect because it occupies a fuzzy middle ground between genuine engagement and outright deflection. We present AsymVerify, a multi-pass verification system that achieves 85% Macro F1 on the SemEval-2026 Task 6 evaluation split (D_{eval}, n=237), ranking second on the official leaderboard. Our core finding is that verification passes targeting opposite failure modes converge to the same result. Passes that upgrade missed commitments and passes that downgrade false acceptances both flow corrections through the middle Ambivalent class, making them functionally interchangeable. This means either pass alone achieves near-optimal performance, enabling flexible deployment. We exploit this property by routing predictions through confidence-gated verification. High-confidence classifications exit early, while uncertain ones receive targeted reexamination calibrated to the likely error direction. This selective approach matches full maj@3 performance at half the inference cost. The method transfers across model families on the development split (D_{dev}, n=308), yielding +5.2 to +21.7 Macro F1 gains over zero-shot baselines using identical prompts.

1 Introduction

Viewers detect political evasion only 39% of the time without explicit cues (Clementson, 2018). Large language models fare worse because their RLHF training rewards charitable interpretation, the precise opposite of what evasion detection requires. Politicians exploit this gap with sophisticated 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) (CLARITY Organizers, 2026). The task uses the QEvasion dataset (Thomas et al., 2024), which 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 ranking. On D_{eval}, AsymVerify reaches 85% Macro F1 (rank #2, tied with CSE-UOI). On D_{dev}, AsymVerify improves zero-shot performance by +5.2 to +21.7 Macro F1 across model families. Confidence analysis on labeled development data shows a clear accuracy gap between high- and low-confidence predictions, motivating 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 addresses a large share of errors.

2 Background

2.1 Task Definition

The CLARITY task (CLARITY Organizers, 2026) 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%), increasing difficulty. Systems are evaluated by Macro F1: $\frac{1}{3}\sum_y 2P_yR_y/(P_y+R_y)$, where P_y and R_y are per-class precision and recall. We report accuracy as a secondary metric.

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

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

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 ranking claims use D_{eval}.

2.2 Related Work

The study of political evasion has a rich history. Bull and Mayer (1993) identified 35 distinct evasion techniques, finding that politicians equivocate on 64% of conflictual questions. 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. Our task differs from stance detection (Mohammad et al., 2016), which classifies opinion direction rather than response clarity; evasion detection requires reasoning about whether commitments were made, not what position was taken.

Our approach builds on self-consistency (Wang et al., 2023), which shows that majority voting over multiple LLM samples improves accuracy by marginalizing over reasoning paths. We apply voting selectively based on confidence, following work on 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. On labeled development data, verbalized confidence separates easier from harder cases, supporting thresholded routing.

Our verification passes can be viewed as a task-specific form of iterative self-refinement (Madaan et al., 2023; Shinn et al., 2023), where reexamination is asymmetrically conditioned on the initial prediction class. Structured chain-of-thought reasoning (Wei et al., 2022) has shown particular promise for implicit subjective classification (Fei et al., 2023), where surface-level features are insufficient; evasion detection shares this property.

Verbalized confidence tends toward overconfidence, though consistency-based aggregation helps mitigate this (Xiong et al., 2024). 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 explicitly counters this tendency through structured taxonomy guidance and skepticism-oriented verification instructions.

Algorithm 1: ASYMVERIFY. Parameters: τ=0.95, k=3. Verification functions $g_{\downarrow}, g_{\uparrow} \in \{0, 1\}$ are LLM calls.

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: majority vote (reuse first call);
$V \leftarrow [\hat{y}]$;
for $i \in 2..k$ **do**
  V.append($f_\theta(q, r).\hat{y}$);
$\hat{y} \leftarrow \text{mode}(V)$;
Pass 2: Downgrade CR/CNR → AMB;
if $\hat{y} \in \{\text{CR}, \text{CNR}\}$ **then**
  if $g_{\downarrow}(q, r, \hat{y}) = 1$ **then**
    $\hat{y} \leftarrow$ AMB;
Pass 3: Upgrade AMB → CR;
if $\hat{y} =$ AMB **then**
  if $g_{\uparrow}(q, r) = 1$ **then**
    $\hat{y} \leftarrow$ CR;
return $\hat{y}$;

3 System Description

AsymVerify operates in up to three passes (Algorithm 1). 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. High-confidence predictions ($c \geq \tau$) exit immediately; low-confidence predictions receive majority voting followed by targeted verification.

The base classifier prompts GPT-5.2 with a structured evasion taxonomy containing nine response subtypes (explicit, implicit, general, partial, dodging, deflection, declining, claims ignorance, clarification). 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. We apply major-

ity voting only when $c < 0.95$: two additional calls are run and the majority label is taken. On D_{dev}, this selective strategy reduces calls by roughly 50% relative to applying maj@3 to all examples.

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. 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. We do not include an AMB→CNR upgrade pass because CNR comprises only 8% of the dataset and CNR→AMB errors account for just 7% of total errors (Table 4), offering minimal recovery potential. Both verification passes run only on low-confidence predictions.

Prefilter for upgrade candidates. Pass 3 is expensive because 67% of predictions are AMB. A rule-based prefilter 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). On D_{dev}, only 13 of 206 AMB predictions pass the prefilter, reducing Pass 3 calls by 94% while preserving 67% prefilter precision on upgrade candidates.

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

Models. Primary: GPT-5.2 with `reasoning_effort=high`. Development-set ablations use GLM-4.7 as the cost-efficient reference model; portability checks use Gemini-3-Flash, GPT-5.1-mini, DeepSeek-V3.2, and Llama-3.3-70B on D_{dev}.

System	Score	Rank
TeleAI	**89.0%**	1
AsymVerify (submission ID 509943)	85.0%	2 (tie)
CSE-UOI	85.0%	2 (tie)
ChatGPT†	51.0%	—

Table 2: Released official leaderboard preview for SemEval-2026 Task 6 (Task 1) on D_{eval} (n=237). AsymVerify is listed at rank #2 with score 0.85 (Macro F1), tied on score with CSE-UOI. †Thomas et al. (2024).

Model	Base F1	+AV F1	Δ
GLM-4.7	55.9%	**73.0%**	+17.1
GPT-5.1-mini	59.0%	64.2%	+5.2
DeepSeek-V3.2	41.0%	62.7%	**+21.7**
Llama-3.3-70B	47.0%	56.7%	+9.7

Table 3: Model portability on D_{dev} (308 examples). All models improve with identical prompts.

Hyperparameters. Confidence threshold $\tau = 0.95$ selected from $\{0.85, 0.90, 0.95, 1.0\}$ on D_{dev}. Temperature 0.1 for base classification, 0.0 for verification. All reported results are single runs; low temperature ensures minimal sampling variance.

5 Results

We first report official leaderboard performance on D_{eval} (Task 1), then turn to development-set analyses on D_{dev} to understand why the approach works.

5.1 Development Analysis (Dev Set)

A natural question is whether AsymVerify's gains are specific to GPT-5.2 or transfer to other models. Table 3 shows that all four models improve by +5 to +22 Macro F1 points over their zero-shot baselines using identical prompts, with the largest gains on the weakest models.

5.2 Cross-Model Class Stability

Improvements are not uniform across classes. Ambivalent detection is highly stable across model families (79.0–81.0 F1), suggesting that the evasion-focused prompt template transfers reliably even when the base model changes. In contrast, Clear Reply varies by 8.2 points (54.8–63.0), 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 re-

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

Table 4: Error patterns on D_{dev} (GLM-4.7, full system, 76 errors). Direct CR↔CNR confusion is rare; errors concentrate at class boundaries.

mains 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, this means model replacement is low risk for AMB-heavy monitoring pipelines, but CNR-sensitive use cases require model-specific prompt adaptation and threshold re-tuning.

The error distribution in Table 4 reveals a striking pattern. 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%), confirming that errors concentrate at the boundaries with AMB rather than between the polar classes.

5.3 Error Coverage by Verification Direction

This boundary concentration has a practical consequence. Over 90% of errors are addressable by the two verification passes. Table 5 reorganizes the same 76 errors by *which pass can repair them*. Pass 2 (downgrade) covers over-acceptance errors where AMB is predicted as CR/CNR (48 errors, 63.2%). Pass 3 (upgrade) covers 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 6), while combining both provides a smaller but consistent additive boost.

The ablation in Table 6 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 +15 points, and combining both yields +17.1, confirming that the passes are complementary rather than redundant. Selective verification achieves this at roughly half the call budget of running all three passes unconditionally (457 versus 924 calls).

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

Table 5: Error coverage by verification direction, derived from Table 4 (76 total errors).

Where do the computational savings come from? Confidence gating routes 162 of 308 examples (53%) directly to output after Pass 1, leaving only 146 for verification (Table 7). The full system uses 457 total calls versus 924 for running all three passes unconditionally, a 50.5% reduction.

Gold labels are now released for Task 1. Our frozen submission (ID 509943) scores 0.85 Macro F1, tied with CSE-UOI at rank #2. Appendix D 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; additional calls can add noise rather than signal. Running maj@3 on high-confidence cases can introduce errors where none existed, while low-confidence predictions benefit from marginalization over reasoning paths. Confidence therefore acts as a useful difficulty proxy, enabling adaptive computation that concentrates resources where they matter.

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 a large share

Configuration	F1	Acc.	Δ
Base (Pass 1)	55.9%	77.6%	—
+ P2 (downgrade)	70.9%	76.0%	+15.0
+ P3 (upgrade)	70.8%	72.1%	+14.9
+ P2 + P3 (full)	**73.0%**	75.3%	**+17.1**

Table 6: Verification pass ablation (GLM-4.7). Both passes contribute independently.

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 7: API call budget on D_{dev} (GLM-4.7, n=308). Confidence gating reduces total calls by 50.5% versus running all passes on every example.

of errors, either path alone addresses a substantial fraction. The rare CR↔CNR confusions (3%) fall outside both paths but are too infrequent to matter.

Even with optimal verification routing, some errors persist. The QEvasion dataset has Fleiss $\kappa = 0.644$ inter-annotator agreement (Thomas et al., 2024), indicating substantial but imperfect human consensus. Remaining errors may partly reflect genuine ambiguity, since some political responses are deliberately crafted to be interpretable as either clear or evasive depending on context.

Appendix C confirms that semantic embeddings alone provide negligible class separation (silhouette = 0.001), motivating prompt-based pragmatic reasoning over retrieval or similarity methods.

The approximately 50% reduction in API calls on D_{dev} halves inference cost without sacrificing quality, and the model-agnostic design ensures the approach remains viable as LLM pricing and capabilities evolve.

The error decomposition also suggests deployment-specific variants. If the priority is minimizing false accusations of "clarity" (high precision on CR/CNR), Pass 2 can be emphasized because it repairs most over-acceptance errors. If the priority is recovering buried commitments (higher CR recall), Pass 3 contributes more directly. The full system remains preferable for leaderboard optimization. However, these one-pass variants provide controllable trade-offs for operational settings with asymmetric risk.

6.1 Persistent Failure Modes

Residual errors cluster into three linguistic regimes that are difficult even for humans. First, *hedged commitments* present a direct answer followed by qualifying rhetoric ("yes, but . . . "), which can be interpreted as either commitment or strategic softening. These cases drive CR→AMB errors when the model over-weights hedging and AMB→CR errors when it over-weights the initial commitment.

Second, *procedural refusals* often appear as discourse management rather than explicit non-answer statements (e.g., rejecting hypotheticals, deferring to another process, or reframing the interviewer premise). These responses can be semantically equivalent to CNR but lack canonical refusal markers ("I will not answer"), which contributes to CNR→AMB errors. The low frequency of this pattern (4 cases in Table 4) suggests that explicit refusals are usually recoverable, but implicit refusals remain brittle.

Third, *conditional commitments* create ambiguity about whether a speaker is committing now or only under future contingencies. In our qualitative examples, statements like "anytime and anyplace . . . we'll be prepared" were sometimes upgraded to CR even when the condition effectively deferred commitment. This failure mode is structurally hard because political language intentionally encodes optionality while preserving the appearance of decisiveness.

These observations suggest concrete prompt refinements. For downgrade verification, stronger tests for condition-triggered non-commitment could reduce false CR labels. For upgrade verification, requiring that the first substantive sentence be both direct *and* unconditional could reduce over-upgrades. More broadly, the remaining error mass appears to come from pragmatic interpretation rather than missing lexical cues. Future gains will likely require richer discourse-level features or calibrated multi-model disagreement signals rather than additional majority voting alone.

7 Conclusion

AsymVerify achieves 85% Macro F1 on SemEval-2026 Task 6 (D_{eval}), ranking second on the official leaderboard. The core finding is that verification passes targeting opposite failure modes converge to the same result, because both flow corrections through the middle Ambivalent class. This convergence property enables confidence-gated routing where high-confidence predictions exit early and uncertain ones receive targeted re-examination, matching full maj@3 performance at half the inference cost. The method transfers across four model families on D_{dev}, yielding +5.2 to +21.7 Macro F1 gains over zero-shot baselines using identical prompts. The remaining errors cluster in linguistically hard cases where even human annotators disagree, suggesting that future progress will require richer discourse-level features or multi-model

disagreement signals rather than additional verification passes.

Limitations

Our confidence threshold and verification prompts were optimized on English-language U.S. political interviews; different political cultures, languages, or interview formats may require re-tuning. While AsymVerify improves all tested models, open-source models still lag frontier models on dev-set Macro F1 (e.g., 56.7–62.7 versus 64.2–73.0 with AsymVerify), suggesting practical deployment on cost-sensitive applications would require model-specific prompt adaptation. Political evasion exists on a spectrum, and the QEvasion dataset has $\kappa = 0.644$ inter-annotator agreement, meaning our system inherits judgment calls where "correct" classification sometimes reflects annotator interpretation rather than ground truth. Finally, we do not evaluate against adversarially-crafted evasions; sophisticated speakers aware of detection systems might develop novel strategies not covered by our taxonomy.

All development-set ablations, error analyses, and cost breakdowns use GLM-4.7 as the reference model. We chose GLM-4.7 deliberately: as the weakest model in our portability set (55.9% base F1), it stress-tests AsymVerify under the least favorable conditions. The portability experiments (Table 3) confirm that gains transfer across four architectures spanning 41–59% base F1, and the submission model (GPT-5.2) achieves the highest absolute score on D_{eval}. However, we did not run per-pass ablations or error decompositions on GPT-5.2 due to the cost of its reasoning-mode API, so we cannot confirm that the specific confidence distributions and verification routing reported in Section 5 hold identically for frontier models.

We also report single-run results with low-temperature decoding and do not include confidence intervals or paired significance tests, so small differences between close variants should be interpreted cautiously. Our verification prompts rely on lexical and discourse cues (e.g., identifying refusal language or the first substantive sentence), which may degrade under noisy transcripts, speech disfluencies, or automatic speech recognition errors. Finally, we classify each question-answer pair independently; incorporating multi-turn context, interviewer follow-ups, and speaker history could resolve cases that are genuinely ambiguous when viewed in isolation.

Ethics Statement

Automated evasion detection could enhance democratic accountability but also enable unfair characterization of political speech. Our system should supplement, not replace, human judgment. Code will be released upon publication.

Acknowledgments

We thank the CLARITY task organizers.

References

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. *Equivocal 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*.

CLARITY Organizers. 2026. SemEval-2026 task 6: CLARITY – detecting political question evasion. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Task description paper.

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

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 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*.

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.

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, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, and 1 others. 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.

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.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 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.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

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 Prompt Templates

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

A.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"}`

A.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"}`

A.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"}`

B Full Model Comparison

Table 8 shows complete results across all models evaluated on the dev set (308 examples).

Model	Method	Acc.	Macro F1	CR F1	AMB F1
Baseline Models (Zero-Shot)					
Grok-4.1-Fast	Zero-shot	78.6%	71.3%	62.4%	84.9%
Gemini-2.5-Flash	Zero-shot	73.4%	64.3%	63.1%	81.5%
Gemini-2.5-Flash-Lite	Zero-shot	77.0%	68.9%	66.7%	82.3%
Gemini-3-Flash	Zero-shot	77.3%	74.8%	66.7%	82.3%
GPT-5.2	Zero-shot	70.2%	70.2%	—	—
Ensemble Baseline (Gemini-3-Flash)					
Ensemble (4 models)	Weighted vote	75.3%	71.2%	68.9%	79.8%
AsymVerify Portability					
GLM-4.7	Zero-shot	77.6%	55.9%	64.0%	83.0%
GLM-4.7	AsymVerify	75.3%	73.0%	63.0%	81.0%
GPT-5.1-mini	Zero-shot	74.0%	59.0%	64.0%	81.0%
GPT-5.1-mini	AsymVerify	73.1%	64.2%	58.0%	79.0%
DeepSeek-V3.2	Zero-shot	66.0%	41.0%	48.0%	75.0%
DeepSeek-V3.2	AsymVerify	72.1%	62.7%	57.0%	80.0%
Llama-3.3-70B	Zero-shot	68.0%	47.0%	37.0%	79.0%
Llama-3.3-70B	AsymVerify	72.4%	56.7%	54.8%	80.7%

Table 8: Full model comparison on dev set (308 examples). GPT-5.2 per-class F1 unavailable due to reasoning-mode API constraints. CNR F1 omitted for space; CNR variation is discussed in Section 5.2.

C Embedding Space Analysis

To verify that evasion detection requires pragmatic reasoning beyond semantic similarity, we project all 3,758 train and test embeddings (Gemini Embedding 001, 3,072 dimensions) into two dimensions using UMAP (McInnes et al., 2018). Figure 1 shows near-total class overlap, with a silhouette score of 0.001 indicating essentially no separation in embedding space. 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 confirms that surface-level semantic features do not distinguish evasive from non-evasive responses, consistent with the failure of contrastive RAG in our early experiments. Evasion detection depends on pragmatic cues such as commitment strength, hedging patterns, and rhetorical structure that general-purpose text embeddings do not capture.

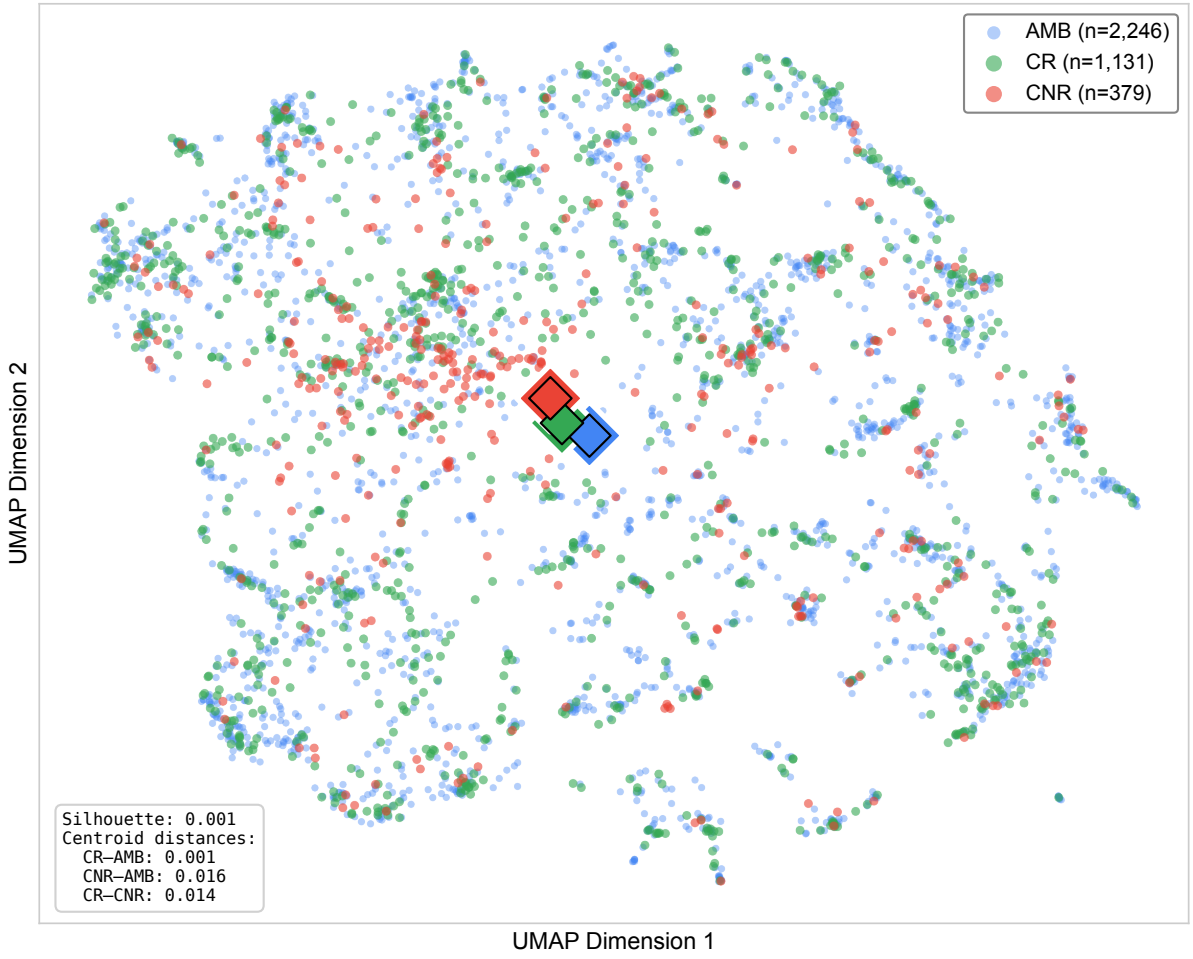

Figure 1: UMAP projection of all 3,758 train and test embeddings (Gemini Embedding 001, 3,072 dimensions). Classes overlap almost entirely (silhouette = 0.001), confirming that evasion detection requires pragmatic reasoning beyond semantic similarity.

D Extended Qualitative Examples

Tables 9 and 10 show representative predictions with full reasoning traces. We select examples where verification *changed* the initial prediction, demonstrating the pipeline's corrective behavior.

Table 9: 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 10: 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.