SELFDOUBT: Uncertainty Quantification for Reasoning LLMs via the Hedge-to-Verify Ratio

Satwik Pandey [*1] **Suresh Raghu** [*1] **Shashwat Pandey** [1]

Abstract

Uncertainty estimation for reasoning language models remains difficult to deploy: sampling-based methods are expensive, while common single-pass proxies such as verbalized confidence or trace length are inconsistent across models. The problem is sharper for proprietary reasoning APIs, which expose neither logits nor reasoning-token probabilities, leaving practitioners with few reliable online signals for failure detection. We propose SELFDOUBT, a single-pass trace diagnostic that extracts behavioral uncertainty signals directly from the reasoning trajectory. Its key signal, the Hedge-to-Verify Ratio (HVR), measures whether expressed doubt is offset by explicit self-checking. Unlike methods that require multiple sampled traces or model internals, SELFDOUBT operates on one observed trace, making it suitable for latency- and cost-constrained deployment. Across seven models and three multi-step reasoning benchmarks (BBH, GPQA-Diamond, and MMLU-Pro), traces with no hedging markers are correct 96% of the time, yielding a high-precision zero-cost accept gate. On the remaining cases, the full SELFDOUBT score significantly outperforms sampling-based Semantic Entropy at 10× lower inference cost. A deployment cascade combining both stages attains 90% accuracy at 71% coverage without task-specific labels. These results position SELFDOUBT as a lightweight trace-level failure diagnostic for reasoning models and a natural building block for future agentic evaluation.

1. Introduction

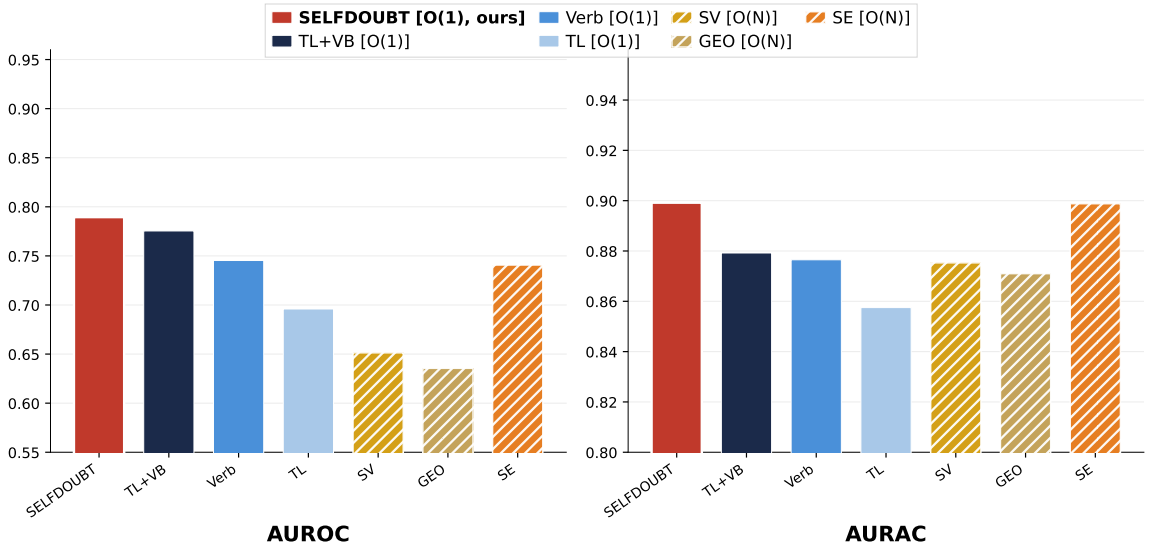
Uncertainty quantification for reasoning language models remains caught between two unappealing extremes. Sampling-based methods, including Semantic Entropy (SE; Kuhn et al., 2023; Farquhar et al., 2024), P(True) self-evaluation (Kadavath et al., 2022), and semantic-geometric approaches (Li et al., 2025; Phillips et al., 2025), require N independent forward passes, which is prohibitive when a single reasoning pass already costs thousands of tokens. Single-pass alternatives exist, but each has a limiting tradeoff: verbalized confidence (Tian et al., 2023; Yoon et al., 2025) is miscalibrated on hard tasks and weaker models; trace length (Devic et al., 2025) correlates with uncertainty only on intermediate-difficulty benchmarks; and probe-based methods (Kossen et al., 2024) require hidden-state access. A further constraint: commercially deployed reasoning APIs do not expose log-probabilities on reasoning tokens, making methods that depend on token-level probabilities (Moslonka et al., 2026; Zhang & Zhang, 2025) architecturally ineligible where UQ is most needed. This setting is closely related to trace-level failure detection in multi-step agents: the trace is observable before the final outcome is known, and any useful diagnostic must be cheap enough to run online.

We propose SELFDOUBT, a single-pass uncertainty framework that extracts behavioral signals directly from the reasoning trace. Its key signal, the Hedge-to-Verify Ratio (HVR), is a single scalar: the number of hedge markers divided by the number of verify markers plus one. Intuitively, hedging expresses doubt (*"maybe," "perhaps," "not sure"*), whereas verification acts on it (*"let me check," "verify," "substitute back"*). HVR therefore captures whether expressed doubt is resolved or left open. Z-score fusion of HVR with verbalized confidence yields the full SELFDOUBT score, requiring zero training, no model internals, and only 90 unlabeled calibration traces per model.

1. **HVR = 0 gate.** We introduce a zero-cost correctness filter requiring only regex matching against a marker dictionary. Traces with zero hedging language are correct 96.1% of the time (25.4% coverage, 0.58% genuine error rate after label-noise correction) across 7 models and 3 datasets.

[*]Equal contribution [1]Independent Researcher, New Delhi, India. Correspondence to: Satwik Pandey <psatwik2711@gmail.com>, Suresh Raghu <sureshraghu0706@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).



(a) Mean AUROC and AURAC across all 21 runs

Figure 1. Headline results. **(a)** Across 21 runs, SELFDOUBT is the only method that leads both metrics at $O(1)$ cost: mean AUROC 0.7895 and mean AURAC 0.8992 (slightly above SE at 0.8988).

2. **SELFDOUBT.** We present an $O(1)$ uncertainty score fusing HVR with verbalized confidence via z-score normalization. SELFDOUBT significantly outperforms Semantic Entropy on discrimination (p=0.001) while matching its selective prediction quality at 10× lower inference cost.

3. **Unsupervised marker discovery.** We introduce an unsupervised per-model pipeline that builds model-specific dictionaries from 90 unlabeled traces: no correctness labels, no manual curation, and no retraining when switching models.

4. **Trace-level failure diagnostic.** Tier 1 (HVR = 0 gate) plus Tier 2 (calibrated z-sum threshold) yields an online accept/defer diagnostic from the trace itself: 71% coverage at 89.7% accuracy, a +9.2 pt lift requiring only 4 stored scalars per model.

2. Related Work

Method landscape: Table 1 organizes UQ methods along the axes critical for reasoning-model deployment: computational cost and method requirements.

Sampling-based UQ: Semantic Entropy (Kuhn et al., 2023; Farquhar et al., 2024) clusters N sampled outputs by meaning; Semantic Volume (Li et al., 2025) and Geometric Uncertainty (Phillips et al., 2025) measure embedding spread; Topo-UQ (Da et al., 2025) computes graph edit distance at $O(N^2)$ cost. All require 10–20× inference cost, which is particularly prohibitive for reasoning models where a single pass already costs thousands of tokens.

Trace-based $O(1)$ methods: Devic et al. (2025) showed trace length is an emergent uncertainty signal, and TL+VB improves over either component alone. Yoon et al. (2025) found reasoning models are better calibrated when verbalizing confidence. The closest concurrent work is Vanhoyweghen et al. (2025), who showed hedging words in CoT traces reduce accuracy by up to 40% relative. Podolak & Verma (2025) further showed that reasoning traces surface self-confidence signals absent in direct-answer generations.

Probe-based and hidden-state methods: Probe-based methods (SEP (Kossen et al., 2024), Latent CoE (Wang et al., 2025)) require hidden states; logprob methods (EPR/WEPR (Moslonka et al., 2026), CoT-UQ (Zhang & Zhang, 2025)) require token probabilities unavailable from production APIs. Both are included in Table 1 but excluded from comparison; What remains missing is an $O(1)$ method that captures deliberation structure from text alone, without trained components or model internals.

Trace diagnostics for agentic failures: Agentic evaluation increasingly emphasizes diagnostics that inspect the trajectory itself rather than only terminal success or failure. SELFDOUBT fits this view by treating the reasoning trace as an observable process signal: hedging exposes expressed uncertainty, while verification exposes attempts to resolve it. The HVR = 0 gate can be interpreted as a zero-cost failure pre-emption rule, and the calibrated SELFDOUBT score provides a lightweight accept/defer diagnostic for the remaining traces. Although our experiments focus on single-episode reasoning benchmarks rather than long-horizon tool-using agents, the same trace-diagnostic framing motivates extension to agentic trajectories.

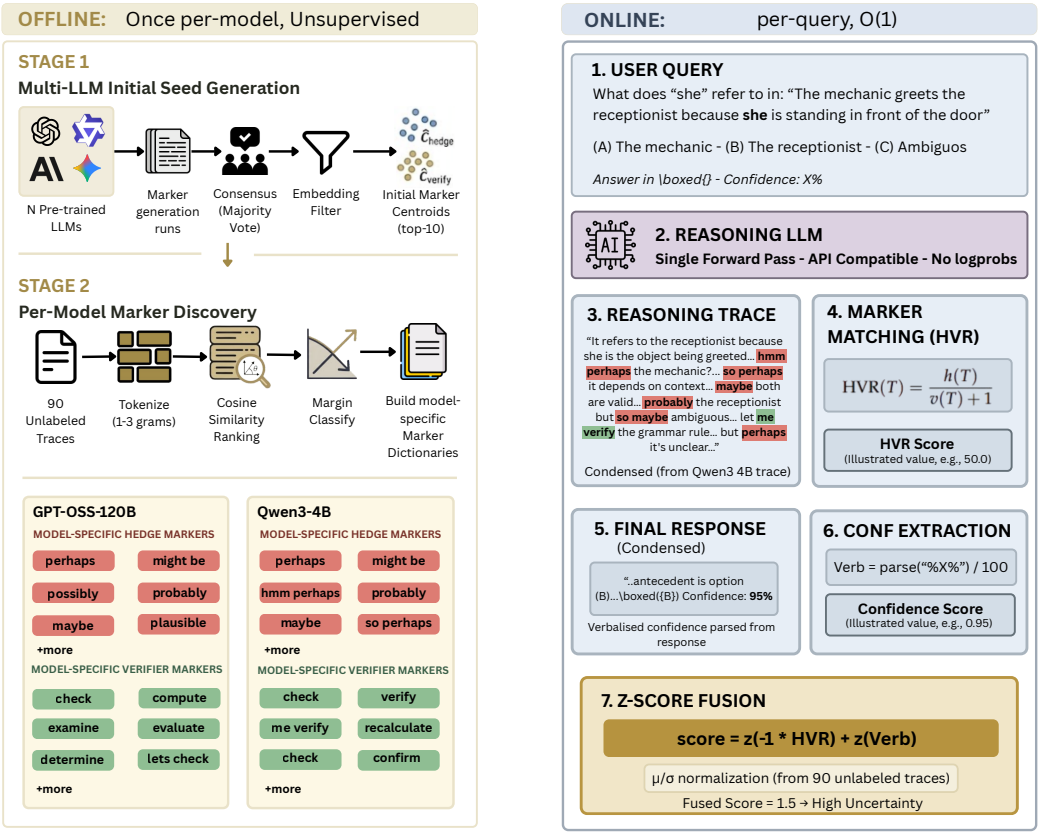


Figure 2. SELFDOUBT pipeline: calibration-time marker discovery and inference-time scoring.

3. SELFDOUBT

Reasoning traces do more than expose a model's final answer: they also expose how the model expresses and resolves doubt during reasoning. SELFDOUBT turns that behavioral signal into a single-pass uncertainty score and trace-level failure diagnostic (Figure 2).

3.1. Data-Driven Marker Discovery

The central challenge in HVR is not the ratio itself, but defining what should count as hedging and verification across different models. Reasoning models do not express uncertainty through a uniform vocabulary, so a fixed lexicon can miss model-specific phrasing and weaken the signal. We therefore use a two-stage data-driven pipeline: Stage 1 builds seed vocabularies for hedging and verification, and Stage 2 expands those seeds into model-specific marker dictionaries from unlabeled traces.

Stage 1: Data-Driven Seed Generation. Rather than hand-specifying initial hedge and verify words, we generate them from model consensus. We query multiple language models to produce candidate single-word lists for two behaviors: language associated with uncertainty, and language associated with actively checking or confirming an answer. Each model is queried five times. We then keep only words that are stable within a model (appearing in a majority of its runs) and stable across models (appearing in a majority of models), yielding a consensus candidate set for each role. Exact prompt templates, model list, and voting rules are given in Appendix M.

To remove weak or noisy candidates, we run an iterative semantic coherence filter using `BAAI/bge-m3` embeddings (Chen et al., 2025); the pipeline is broadly robust to embedding model choice (Appendix E). We embed all candidate words, compute a centroid for the active set, drop outliers below a cosine-similarity threshold of 0.7 (while preserving at least 10 words), recompute the centroid, and repeat for up to 6 rounds. The surviving words are ranked by cosine-to-centroid coherence and materialized into subsets $\{\text{top_2}, \ldots, \text{top_20}\}$. We use `top_10` as the default seed set, selected by the seed-size ablation (Appendix F).

Stage 2: Per-Model Marker Expansion.

Given the seed vocabularies from Stage 1, we expand them into model-specific marker dictionaries using unlabeled reasoning traces. We sample 90 traces per model across the

Table 1. Uncertainty quantification methods for language models. $O(1)$ methods above the first rule; $O(N)$ and $O(N^2)$ below.

Method	Cost	Requirements
Verb. Conf. (Tian et al., 2023; Yoon et al., 2025)	$O(1)$	Prompt mod.
Trace Length (Devic et al., 2025)	$O(1)$	Trace
TL+VB (Devic et al., 2025)	$O(1)$	Trace + prompt mod.
Lex. Hints (Vanhoyweghen et al., 2025)	$O(1)$	Trace + sentiment scorer
CoT-UQ (Zhang & Zhang, 2025)	$O(1)$	Token logits
EPR/WEPR (Moslonka et al., 2026)	$O(1)$	Top-K logprobs
SEP (Kossen et al., 2024)	$O(1)$	Hidden states
Latent CoE (Wang et al., 2025)	$O(1)$	Hidden states
P(True) self-eval. (Kadavath et al., 2022)	$O(N)$	Self-eval + sampled refs (often logprobs)
SE (Kuhn et al., 2023; Farquhar et al., 2024)	$O(N)$	N samples + NLI
SV (Li et al., 2025)	$O(N)$	N samples + log-det(Gram)
GEO (Phillips et al., 2025)	$O(N)$	N samples + convex hull
Topo-UQ (Da et al., 2025)	$O(N^2)$	N samples + GED
SELFDOUBT	$\mathbf{O(1)}$	**Trace + Prompt mod.**

evaluation datasets and extract candidate 1–3-grams from their text. To suppress idiosyncratic phrasing, we retain only n-grams that appear in at least 9 out of the 90 distinct traces (10%). Appendix D shows that marker discovery is broadly stable across calibration set sizes, so we use 90 traces as our standard setting.

We then embed each candidate with `bge-m3` and score it by how much more closely it aligns with the verify centroid than with the hedge centroid:

$$\Delta(g) = \cos(g,\, c_{\text{verify}}) - \cos(g,\, c_{\text{hedge}})$$

A candidate enters the verify dictionary if $\Delta(g) > \tau_{\text{verify}}$, enters the hedge dictionary if $\Delta(g) < -\tau_{\text{hedge}}$, and is otherwise discarded. This creates a model-specific neutral band around zero that filters ambiguous or weakly aligned n-grams. The verify and hedge thresholds are set separately for each model so that semantically relevant markers are retained while domain-specific vocabulary is excluded. Exact threshold values are given in Appendix B; sensitivity to these choices is analyzed in Appendix N.

The result is a per-model hedge dictionary $\mathcal{H}$ and verify dictionary $\mathcal{V}$, compiled into word-boundary regex patterns for trace scanning at inference time. Markers transfer across datasets (Appendix G).

3.2. Hedge-to-Verify Ratio

Given the per-model dictionaries $\mathcal{H}$ and $\mathcal{V}$ produced above, we define the signal that consumes them. Let $h(T)$ be the total number of hedge-marker occurrences in trace T (matched via word-boundary regex against $\mathcal{H}$, non-overlapping, longest-match-first), and $v(T)$ the corresponding count against $\mathcal{V}$. The Hedge-to-Verify Ratio is:

$$\text{HVR}(T) = \frac{h(T)}{v(T) + 1} \tag{1}$$

The $+1$ in the denominator ensures finite values when no verify markers appear. We orient HVR so that larger values correspond to greater uncertainty; empirically, this direction is consistent across all 21 runs. The distribution is zero-inflated by construction: any trace with zero hedge matches yields $\text{HVR} = 0$ exactly. We analyze this HVR$= 0$ regime separately in Section 5 and later operationalize it as a deployment-time accept rule in Section 6.

Stage 2 per-model marker expansion is retained as the default because the HVR$= 0$ gate is sensitive to marker coverage (Appendix H).

3.3. Verbalized Confidence

In addition to trace-derived uncertainty, we use a direct self-reported signal: *verbalized confidence*. For each query, the model is prompted to output a final answer together with a confidence in percent form. We parse this value from the final answer region and map it to a scalar $V \in [0, 1]$ by dividing the reported percentage by 100, where larger V indicates greater reported confidence in correctness.

3.4. SELFDOUBT Score

HVR and verbalized confidence capture complementary failure modes. HVR captures uncertainty revealed through deliberation behavior, while verbalized confidence captures uncertainty the model explicitly reports when asked. We combine them with additive z-score fusion:

$$s_{\text{SD}}(T) = z_r(-\text{HVR}(T)) + z_r(V), \tag{2}$$

where V is the verbalized confidence score and $z_r(x) = (x - \mu_r)/\sigma_r$ standardizes each channel within run r over the usable joined subset. The negation aligns HVR with correctness, and standardization makes the two channels

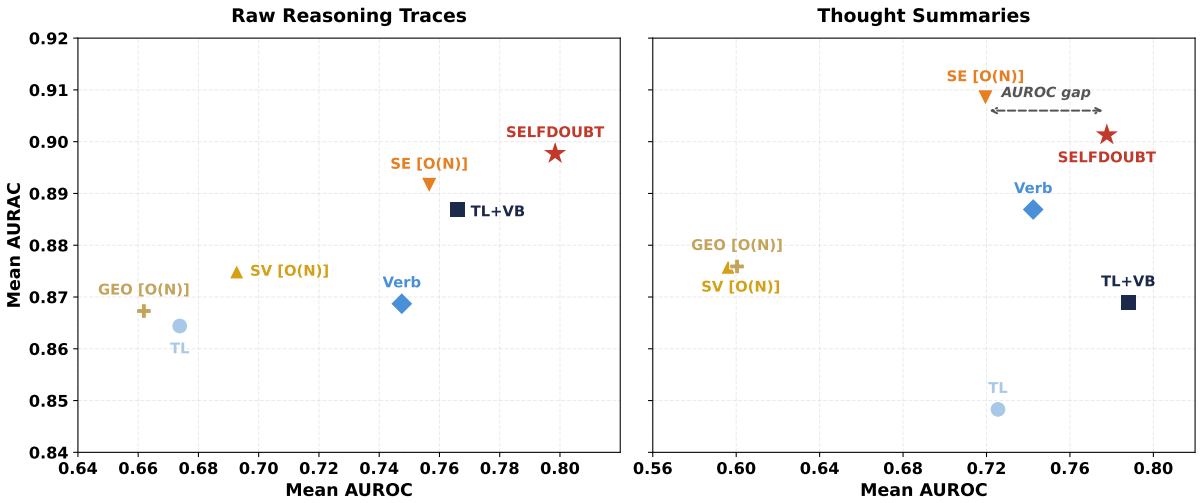


Figure 3. AUROC–AURAC consistency on thought summary models. SELFDOUBT occupies the upper-right quadrant among $O(1)$ methods. SE shows strong AURAC but weak AUROC on compressed traces.

commensurable despite model-specific differences in scale.

4. Experimental Setup

Models. We evaluate seven reasoning models spanning two trace types. Four produce full reasoning traces via local inference: Qwen3 4B, Qwen3 14B, GPT OSS 20B, and GPT OSS 120B. Three produce compressed thought summaries via commercial APIs: Claude Sonnet 4.6, Grok 4.1 Fast, and Gemini 2.5 Flash. Thought summaries expose a provider-compressed version of internal reasoning rather than the full token-level trace, making uncertainty estimation harder (Section 5.4). Model details are in the respective technical reports (OpenAI et al., 2025; Yang et al., 2025; Anthropic, 2026; xAI, 2025; Gemini Team, 2025).

Datasets. GPQA-Diamond (198 questions, graduate-level science; Rein et al., 2023), BBH (Big-Bench Hard) (300 stratified questions, diverse reasoning; Suzgun et al., 2022), and MMLU-Pro (300 stratified questions, 10-option multiple choice; Wang et al., 2024). Together these span three difficulty profiles and question styles. All datasets are multiple-choice. We evaluated 21 runs: 7 models $\times$ 3 datasets.

Baselines. All baselines are evaluated on the same samples. $O(1)$ baselines: Verbalized Confidence (Verb; Tian et al., 2023; Yoon et al., 2025), Trace Length (TL; Devic et al., 2025), and TL+VB (Devic et al., 2025). $O(N)$ baselines: Semantic Entropy (SE; Farquhar et al., 2024), Semantic Volume (SV; Li et al., 2025), and Geometric Uncertainty (GEO; Phillips et al., 2025). We run all sampling-based baselines with N=10 samples, incurring roughly 10$\times$ the generation cost of a single-pass approach.

Metrics. *AUROC* (primary): area under the ROC curve measuring discrimination, i.e., how well the score separates correct from incorrect answers. *AURAC* (secondary): area under the Risk–Coverage curve measuring selective prediction quality, i.e., how well the method's ranking supports deferral policies across all possible thresholds.

Statistical testing. We use paired Wilcoxon signed-rank tests (one-sided alternative: SELFDOUBT $>$ baseline) on 21 matched model$\times$dataset run pairs.

5. Results

5.1. The Zero-Hedge Regime

Pooled across 21 runs (5455 queries), traces with $\mathrm{HVR} = 0$ are correct 96.1% of the time (1330/1384) at 25.4% coverage (Table 2). The property holds at $\geq$90% precision for 6 of 7 models; Gemini is the sole outlier, producing only 7 zero-hedge traces across all three datasets.

Manual error analysis of the 54 raw disagreements reveals that only 8/1384 (0.58%) represent genuine model failures: the model answered with high confidence and was factually wrong. The remaining 46 disagreements are grading artifacts (30 cases: incorrect answer keys or format mismatches) or ambiguous question labels (16 cases). The label-noise-corrected precision is 99.4% (Wilson 95% CI: [98.9%, 99.7%]). Details in Appendix C.

Thought summary models have higher coverage (35.4%) than raw-trace models (17.7%), reflecting compressed traces that surface fewer hedge markers. Coverage is highly model-dependent, ranging from 0.9% (Gemini) to 53.3% (Claude).

5.2. Main Comparison

Moving from the binary regime to the continuous score, Table 3 reports mean AUROC and AURAC across all 21

Table 2. Precision and coverage of the HVR = 0 subset with 95% Wilson confidence intervals, grouped by trace format. Counts are shown as N_0/N.

Group	Model	N_0/N	Coverage	Accuracy (HVR = 0)	Wilson 95% CI
Full traces	Qwen 3 4B	31/722	4.3%	90.3%	[75.1%, 96.7%]
	Qwen 3 14B	77/794	9.7%	96.1%	[89.2%, 98.7%]
	GPT OSS 20B	233/771	30.2%	96.1%	[92.8%, 98.0%]
	GPT OSS 120B	205/798	25.7%	96.6%	[93.1%, 98.3%]
Thought summaries	Claude Sonnet 4.6	425/797	53.3%	98.1%	[96.3%, 99.0%]
	Grok 4.1 Fast	406/798	50.9%	94.8%	[92.2%, 96.6%]
	Gemini 2.5 Flash	7/775	0.9%	57.1%	[25.1%, 84.2%]
All	**Pooled (21 runs)**	**1384/5455**	**25.4%**	**96.1%**	[94.9%, 97.0%]

runs, grouped by trace format.

SELFDOUBT achieves the highest mean AUROC (0.7895) and mean AURAC (0.8992) among all methods at any cost. It is the only $O(1)$ method competitive with SE on AURAC, matching SE's 0.8988 at a tenth of the inference cost. Per-dataset, SELFDOUBT is strongest on BBH (4/7 AUROC wins), while GPQA and MMLU-Pro show tighter margins and more competitive baseline behavior (Appendix A).

SELFDOUBT's primary advantage is not per-run dominance but cross-model, cross-metric, cross-trace-type consistency: best mean on both AUROC and AURAC at $O(1)$ cost, with the smallest variance across trace formats. This consistency is what makes it deployable. Section 6 evaluates this directly.

5.3. Statistical Significance

Table 4 reports paired significance tests across 21 matched runs. SELFDOUBT significantly outperforms SE on AUROC ($p = 0.001$), confirming that an $O(1)$ method can beat the $O(N)$ gold standard at discrimination. The AURAC comparison is not significant ($p = 0.069$), indicating equivalent selective prediction quality at one-tenth the cost. SELFDOUBT also significantly outperforms Verb-only on both metrics ($p < 0.001$), confirming that HVR carries signal beyond verbalized confidence alone. Against TL+VB, results are not significant in either direction. On the full-trace subgroup (12 runs), SELFDOUBT significantly beats TL+VB on both metrics ($p = 0.026$ AUROC, $p = 0.039$ AURAC); on thought summaries, results are competitive but not significant (Appendix J).

5.4. Where SELFDOUBT Loses

SELFDOUBT does not achieve the best AUROC on the majority of individual runs. Two failure modes account for most losses. First, when verbalized confidence is weak on a run (notably Qwen3 4B), adding Verb dilutes HVR signal. HVR alone outperforms the fused score on these runs. Second, Gemini's compressed thought summaries produce very sparse n-gram candidates (0.9% coverage), making HVR nearly constant and giving trace length a structural advantage on that model. This suggests a minimum trace richness below which SELFDOUBT cannot operate reliably.

6. Deployment Policy

A deployable uncertainty method must do more than rank responses: for each query, it must decide whether to answer or defer before an erroneous response is accepted or propagated downstream. We operationalize SELFDOUBT as a two-stage cascade. Stage 1 auto-accepts traces with zero hedging language (the HVR = 0 gate). Stage 2 applies calibrated z-score fusion to the remainder, deferring queries below a tunable threshold.

Runtime rule. Deployment uses the per-model marker dictionaries from Section 3 together with per-model calibration means and standard deviations for HVR and verbalized confidence, estimated from the 90-trace calibration set.

Tier 1 (HVR = 0 gate). The first stage exploits the sharp structural effect identified in Section 5.1: traces with no hedging language form a distinct high-precision subset. Operationally, these cases can be accepted immediately, since detecting them requires only marker matching and no score computation. To avoid unstable behavior on models where zero-hedge traces are too rare to calibrate reliably, we enable the gate only when the calibration set contains at least four such examples ($N_0 \geq 4$); in practice this excludes only Gemini (N_0=2). Under this deployed policy, Tier 1 alone answers 25.2% of queries at 96.3% accuracy. (99.4% after label-noise correction; Section 5.1)

Tier 2 (calibrated z-sum). The remaining queries (HVR $>$ 0) are scored by the calibrated fusion

$$s = z_{\text{cal}}(-\text{HVR}) + z_{\text{cal}}(V) = \frac{-\text{HVR} - \mu_{\text{hvr}}}{\sigma_{\text{hvr}}} + \frac{V - \mu_{\text{v}}}{\sigma_{\text{v}}},$$

where $\mu_{\text{hvr}}, \sigma_{\text{hvr}}, \mu_{\text{v}}, \sigma_{\text{v}}$ are estimated from the calibration set, and the query is accepted when $s \geq \tau$. Because both terms are standardized on that sample, τ=0 is the natural

Table 3. Main results across 21 runs (7 models × 3 datasets). "All" is the mean over all runs; "Trace" and "Summ." are means over the 12 full-trace and 9 thought-summary runs, respectively. $O(1)$ methods above rule, $O(N)$ below. **Bold**: best; underline: second-best.

Method	Cost	All		Trace		Summ.	
		AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
SELFDOUBT (ours)	$O(1)$	**0.7895**	**0.8992**	**0.7984**	**0.8977**	0.7777	0.9013
TL+VB	$O(1)$	0.7754	0.8792	0.7659	0.8869	**0.7880**	0.8689
HVR (ours)	$O(1)$	0.7509	0.8883	0.7674	0.8884	0.7288	0.8882
Verb	$O(1)$	0.7453	0.8765	0.7475	0.8687	0.7424	0.8869
TL	$O(1)$	0.6959	0.8575	0.6738	0.8644	0.7255	0.8483
SE	$O(N)$	0.7407	0.8988	0.7566	0.8916	0.7195	**0.9085**
SV	$O(N)$	0.6513	0.8753	0.6927	0.8749	0.5961	0.8758
GEO	$O(N)$	0.6355	0.8710	0.6619	0.8673	0.6004	0.8759

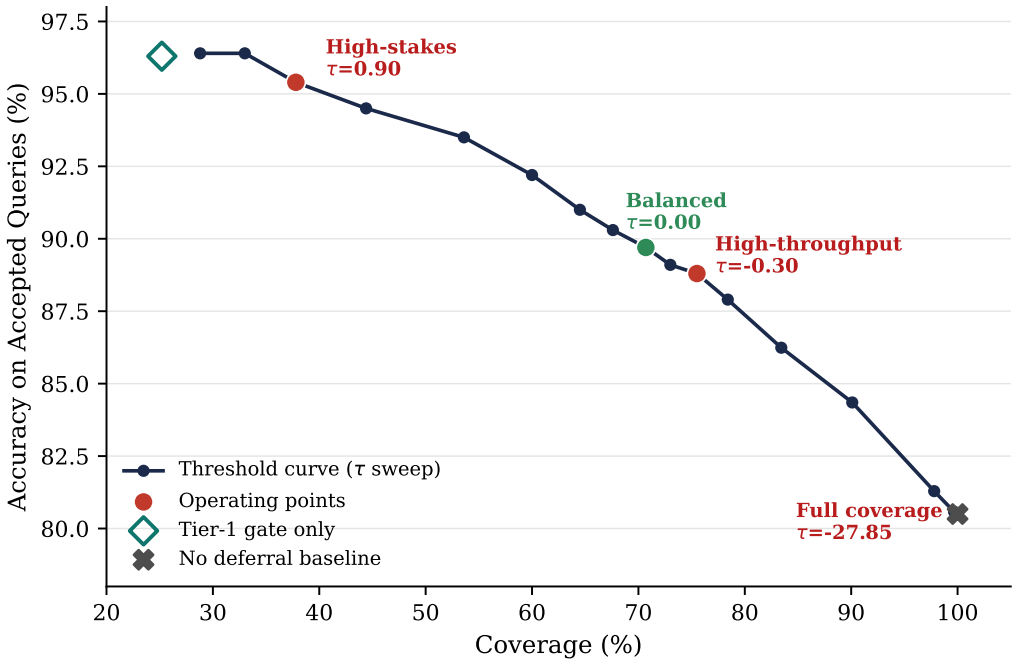


Figure 4. Accuracy–coverage curve for the SELFDOUBT deployment cascade.

Table 4. Paired Wilcoxon signed-rank tests (one-sided: SELFDOUBT > baseline; 21 matched run pairs). W–D–L = wins–draws–losses for SELFDOUBT.

Comparison	Metric	Mean Δ	W–D–L	W	p (one-sided)	Sig.?
vs. SE	AUROC	+0.049	17–0–4	198	0.001	**Yes**
vs. SE	AURAC	+0.000	16–0–5	159	0.069	No
vs. Verb	AUROC	+0.044	16–1–4	190	<0.001	**Yes**
vs. Verb	AURAC	+0.023	18–0–3	223	<0.001	**Yes**
vs. TL+VB	AUROC	+0.014	12–0–9	146	0.152	No
vs. TL+VB	AURAC	+0.020	11–0–10	142	0.187	No

symmetric default: it accepts queries whose fused evidence is above the calibration mean and defers the rest. More conservative or more permissive operating points are obtained by shifting τ upward or downward according to domain risk tolerance.

Calibration cost. Moving from oracle normalization (evaluation-set statistics, unavailable at deployment time) to calibration-set normalization (90-trace statistics) costs 0.48 percentage points of AURAC (0.8992 → 0.8944). This small gap shows that z-score fusion transfers from evaluation-time normalization to deployment-time calibration with minimal loss.

Figure 4 shows the full accuracy–coverage tradeoff. The curve is smooth and monotonic, confirming that practitioners can tune τ to any desired operating point. At τ=0, the cascade reaches the balanced operating point emphasized in the paper: 70.7% coverage at 89.7% accuracy. Higher thresholds push the system toward high-stakes use cases; lower thresholds recover more throughput while degrading gracefully toward the full-coverage baseline.

The calibrated z-sum remains the strongest Tier-2 ranker we

tested: holding Tier 1 fixed, it achieves cascade AURAC 0.8944 and Tier-2 AUROC 0.7572 on the $\mathrm{HVR} > 0$ subset, compared with 0.8830 and 0.7250 for SE Appendix I. Deployment requires only per-model marker dictionaries and four stored scalars, all computed from the same 90 unlabeled traces used for marker discovery.

7. Limitations

1. **HVR = 0 coverage is model-dependent.** For models with very sparse thought summaries, the Tier-1 gate provides negligible throughput benefit, and the practical value of the deployment cascade is reduced.
2. **Style sensitivity.** HVR is defined over surface-form hedging and verification markers. A model or prompt that suppresses hedging language, or that inserts verification-like phrases without substantive checking, could distort both the continuous HVR score and the HVR = 0 gate. Robustness to such stylistic drift is untested.
3. **No uncertainty decomposition.** SELFDOUBT produces a single uncertainty score and does not distinguish epistemic from aleatoric sources of uncertainty. This is sufficient for selective prediction, which requires only a ranking, but limits applicability in settings where the source of uncertainty matters.
4. **Not yet evaluated on long-horizon agents.** Our experiments study single-episode reasoning traces rather than tool-using agents with memory, environment feedback, or planner resets. Extending SELFDOUBT to long-horizon agent trajectories, where early failures can silently contaminate later steps, is a natural next step but remains untested.
5. **Multiple-choice only.** All three evaluation datasets use multiple-choice format. Whether HVR and SELFDOUBT generalize to free-form generation tasks (open-ended QA, code generation, mathematics with step-by-step scoring) is untested.

8. Conclusion

Two findings from our evaluation stand out.

First, HVR = 0 defines a high-precision correctness gate: pooled across 21 runs, zero-hedge traces are correct 96.1% of the time at 25.4% coverage, with only 0.58% genuine errors after label-noise correction. Second, z-score fusion of HVR with verbalized confidence achieves the best mean AUROC and AURAC among all $O(1)$ methods. It significantly outperforms the $O(N)$ baseline Semantic Entropy on discrimination (p=0.001) while matching it on selective prediction at one-tenth the inference cost. A production cascade built from these components reaches 71% coverage at 89.7% accuracy, a 9.2-point lift over the no-deferral baseline, requiring no task-specific labels and only 90 unlabeled calibration traces per model.

Beyond deployment, these results point to a stable textual regularity in reasoning traces: correctness correlates with how models hedge and self-check, a pattern consistent across seven models, two trace types, and three datasets. Whether other trace features, such as revision patterns, confidence trajectories, self-correction timing, or tool-call recovery behavior, carry additional signal is an open question as reasoning models are deployed in agentic and high-stakes settings.

Reproducibility Statement. SELFDOUBT requires no model training, fine-tuning, or access to model internals. The full pipeline runs from 90 unlabeled reasoning traces per model using only regex matching and arithmetic. All marker dictionaries, per-model thresholds, and Stage 2 expansion parameters are provided in Appendices B and M. Prompts for verbalized confidence extraction and Stage 1 seed generation are reproduced verbatim in Appendix M. Deployment requires four stored scalars per model (two means, two standard deviations) estimated from the same 90-trace calibration set.

Impact Statement

This paper presents work whose goal is to improve uncertainty estimation for reasoning language models, especially in settings where users must decide when to trust, defer, or further verify model outputs. Better single-pass uncertainty signals may reduce overreliance on incorrect model answers and lower the cost of deploying selective prediction systems. At the same time, behavioral uncertainty signals such as hedging and verification language can be style-sensitive and should not be treated as guarantees of correctness, particularly in high-stakes domains. We therefore view SELFDOUBT as a decision-support tool for prioritizing review or deferral, not as a replacement for human oversight or domain-specific validation.

References

Anthropic. Claude sonnet 4.6 system card. Anthropic system card, 2026. URL https://anthropic.com/claude-sonnet-4-6-system-card. Accessed: 2026-03-24.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. 2025. URL https://arxiv.org/abs/2402.03216.

Da, L., Liu, X., Dai, J., Cheng, L., Wang, Y., and Wei, H. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. 2025. URL https://arxiv.org/abs/2502.17026.

Devic, S., Peale, C., Bradley, A., Williamson, S., Nakkiran, P., and Gollakota, A. Trace length is a simple uncertainty signal in reasoning models. 2025. URL https://arxiv.org/abs/2510.10409.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025. URL https://arxiv.org/abs/2507.06261.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. 2022. URL https://arxiv.org/abs/2207.05221.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms. 2024. URL https://arxiv.org/abs/2406.15927.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. 2023. URL https://arxiv.org/abs/2302.09664.

Li, X., Yu, Z., Zhang, Z., Zhuang, Y., Shah, S., Sadagopan, N., and Beniwal, A. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms. 2025. URL https://arxiv.org/abs/2502.21239.

Moslonka, C., Randrianarivo, H., Garnier, A., and Malherbe, E. Learned hallucination detection in black-box llms using token-level entropy production rate. 2026. URL https://arxiv.org/abs/2509.04492.

OpenAI, :, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., and Zhao, S. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenRouter. Api parameters. https://openrouter.ai/docs/api/reference/parameters, a. OpenRouter documentation. Accessed 2026-03-28.

OpenRouter. Provider routing. https://openrouter.ai/docs/guides/routing/provider-selection, b. OpenRouter documentation. Accessed 2026-03-28.

OpenRouter. Quickstart. https://openrouter.ai/docs/quickstart, c. OpenRouter documentation. Accessed 2026-03-28.

Phillips, E., Wu, S., Molaei, S., Belgrave, D., Thakur, A., and Clifton, D. Geometric uncertainty for detecting and correcting hallucinations in llms. 2025. URL https://arxiv.org/abs/2509.13813.

Podolak, J. and Verma, R. Read your own mind: Reasoning helps surface self-confidence signals in llms. 2025. URL https://arxiv.org/abs/2505.23845.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. 2023. URL https://arxiv.org/abs/2311.12022.

Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. 2024. URL https://arxiv.org/abs/2310.16049.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. 2022. URL https://arxiv.org/abs/2210.09261.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. 2023. URL https://arxiv.org/abs/2305.14975.

Vanhoyweghen, A., Verbeken, B., Algaba, A., and Ginis, V. Lexical hints of accuracy in llm reasoning chains. 2025. URL https://arxiv.org/abs/2508.15842.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. 2024. URL https://arxiv.org/abs/2406.01574.

Wang, Y., Zhang, P., Yang, B., Wong, D. F., and Wang, R. Latent space chain-of-embedding enables output-free llm self-evaluation. 2025. URL https://arxiv.org/abs/2410.13640.

xAI. Grok 4.1 model card. xAI model card, 2025. URL https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf. Accessed: 2026-03-24.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. 2025. URL https://arxiv.org/abs/2505.09388.

Yoon, D., Kim, S., Yang, S., Kim, S., Kim, S., Kim, Y., Choi, E., Kim, Y., and Seo, M. Reasoning models better express their confidence. 2025. URL https://arxiv.org/abs/2505.14489.

Zhang, B. and Zhang, R. Cot-uq: Improving response-wise uncertainty quantification in llms with chain-of-thought. 2025. URL https://arxiv.org/abs/2502.17214.

Appendix A: Full Per-Run Results

Table 5 and Table 6 report per-run AUROC for all 21 runs. Tables 7 and 8 report per-run AURAC. Bold indicates per-run SOTA. SELFDOUBT achieves the highest or second-highest AUROC on the majority of raw-trace runs. On thought summaries, per-run leadership is split between SELFDOUBT, TL+VB, and SE, consistent with the non-significant pairwise comparisons reported in Table 4.

Table 5. Per-run AUROC: raw reasoning trace models.

Dataset	Model	N	SD	HVR	Verb	TL	TL+VB	SE	SV	GEO
BBH	gpt_120b	300	**0.832**	0.742	**0.832**	0.623	0.817	0.779	0.749	0.729
BBH	gpt_20b	292	0.773	0.703	0.778	0.568	0.780	**0.819**	0.776	0.784
BBH	qwen3_4b	300	**0.747**	0.733	0.667	0.649	0.702	0.732	0.628	0.505
BBH	qwen3_14b	300	**0.799**	0.746	0.785	0.524	0.723	0.727	0.537	0.611
GPQA	gpt_120b	198	**0.844**	0.813	0.835	0.761	0.839	0.760	0.758	0.697
GPQA	gpt_20b	180	0.736	0.721	0.727	0.767	**0.777**	0.646	0.689	0.556
GPQA	qwen3_4b	197	0.808	**0.812**	0.633	0.777	0.770	0.751	0.711	0.684
GPQA	qwen3_14b	198	0.736	0.707	0.655	0.736	**0.752**	0.733	0.672	0.613
MMLU	gpt_120b	300	0.847	0.804	**0.848**	0.696	0.832	0.756	0.758	0.749
MMLU	gpt_20b	299	0.864	0.814	0.826	0.685	0.818	**0.873**	0.804	0.711
MMLU	qwen3_4b	225	0.768	**0.802**	0.639	0.667	0.655	0.707	0.666	0.643
MMLU	qwen3_14b	296	**0.828**	0.814	0.747	0.631	0.728	0.797	0.566	0.662

Table 6. Per-run AUROC: thought summary models.

Dataset	Model	N	SD	HVR	Verb	TL	TL+VB	SE	SV	GEO
BBH	claude	300	**0.809**	0.761	0.795	0.611	0.714	0.653	0.570	0.535
BBH	grok	300	0.826	0.771	0.821	0.723	**0.838**	0.785	0.563	0.700
BBH	gemini	293	0.640	0.645	0.532	**0.827**	0.770	0.663	0.516	0.614
GPQA	claude	198	0.787	0.686	0.807	0.772	**0.819**	0.679	0.564	0.503
GPQA	grok	198	0.831	0.791	0.801	0.809	**0.887**	0.709	0.635	0.687
GPQA	gemini	197	0.662	0.638	0.592	0.635	0.666	**0.771**	0.649	0.609
MMLU	claude	299	0.911	0.840	**0.918**	0.793	0.903	0.750	0.670	0.640
MMLU	grok	300	0.774	0.720	0.771	0.749	**0.791**	0.710	0.590	0.519
MMLU	gemini	285	**0.759**	0.707	0.645	0.612	0.704	0.756	0.608	0.596

Table 7. Per-run AURAC: raw reasoning trace models.

Dataset	Model	N	SD	HVR	Verb	TL	TL+VB	SE	SV	GEO
BBH	gpt_120b	300	**0.975**	0.947	0.962	0.932	0.971	0.955	0.972	0.970
BBH	gpt_20b	292	0.963	0.955	0.948	0.933	**0.966**	0.947	0.945	0.946
BBH	qwen3_4b	300	0.933	0.933	0.921	0.912	0.920	0.933	**0.939**	0.908
BBH	qwen3_14b	300	**0.965**	0.949	0.954	0.902	0.949	0.918	0.908	0.929
GPQA	gpt_120b	198	**0.882**	0.871	0.877	0.834	0.879	0.878	0.881	0.864
GPQA	gpt_20b	180	0.715	0.716	0.681	0.726	0.731	0.790	**0.821**	0.779
GPQA	qwen3_4b	197	**0.860**	0.859	0.749	0.835	0.836	0.795	0.794	0.783
GPQA	qwen3_14b	198	0.813	0.798	0.738	0.803	**0.813**	0.810	0.746	0.730
MMLU	gpt_120b	300	0.935	0.917	**0.938**	0.895	0.935	0.924	0.924	0.922
MMLU	gpt_20b	299	**0.941**	0.924	0.931	0.889	0.929	0.929	0.884	0.862
MMLU	qwen3_4b	225	0.862	0.872	0.828	0.834	0.813	**0.892**	0.839	0.832
MMLU	qwen3_14b	296	**0.930**	0.920	0.896	0.878	0.901	0.928	0.845	0.883

AURAC is tighter across methods, with per-run margins often below 1 percentage point. SELFDOUBT leads most frequently on raw traces; on thought summaries, TL+VB is the most common per-run leader.

Table 8. Per-run AURAC: thought summary models.

Dataset	Model	N	SD	HVR	Verb	TL	TL+VB	SE	SV	GEO
BBH	claude	300	0.953	0.948	0.949	0.955	**0.967**	0.916	0.941	0.938
BBH	grok	300	0.970	0.960	0.964	0.957	**0.972**	0.934	0.902	0.945
BBH	gemini	293	0.709	0.711	0.634	0.371	0.371	0.904	0.906	**0.936**
GPQA	claude	198	0.925	0.881	0.928	0.913	**0.936**	0.916	0.905	0.876
GPQA	grok	198	0.947	0.937	0.938	0.925	**0.960**	0.929	0.772	0.908
GPQA	gemini	197	0.790	0.785	0.771	0.785	0.792	**0.864**	0.820	0.803
MMLU	claude	299	0.977	0.966	**0.978**	0.954	0.976	0.898	0.882	0.871
MMLU	grok	300	0.900	0.878	0.899	0.904	**0.917**	0.884	0.853	0.780
MMLU	gemini	285	**0.940**	0.928	0.921	0.871	0.930	0.931	0.902	0.825

Appendix B: Marker Discovery Details

Stage 1 centroid filtering parameters. Embedding model: `BAAI/bge-m3` (Chen et al., 2025); similarity threshold: 0.7; minimum keep count: 10; maximum iterations: 6. Observed: hedge candidates remained stable across iterations (no drops); verifier candidates had noisy outliers removed (e.g., *"hmm," "plugging"*). Thresholds were selected per model by inspecting margin distributions and choosing cutoffs that preserve semantically coherent markers while excluding domain-specific vocabulary; sensitivity to this choice is reported in Appendix N.

Table 9. Per-model Stage 2 thresholds for marker classification.

Model	τ_{verify}	τ_{hedge}
Claude Sonnet 4.6	0.09	0.08
Gemini 2.5 Flash	0.12	0.05
`gpt-oss-20b`	0.14	0.20
`gpt-oss-120b`	0.10	0.15
Grok 4.1 Fast	0.10	0.10
Qwen3	0.17	0.20
Qwen3-14B	0.15	0.15

Per-model Stage 2 thresholds.

Appendix C: Audit of HVR=0 Raw Disagreements

This appendix details the manual audit underlying the HVR = 0 analysis. Across the 21 pooled runs (7 models $\times$ 3 datasets), the evaluation set contains 5455 total rows, of which 1384 have $\text{HVR} = 0$ (25.4% coverage). Under the original benchmark labels, 1330 of these 1384 rows are marked correct, giving a raw precision of 96.10% and leaving 54 apparent false accepts for audit.

What was audited. We manually re-examined *all 54* HVR = 0 rows that were marked incorrect by the original benchmark labels. The goal of this audit was not to relabel the entire benchmark, but to determine whether each apparent HVR = 0 failure represented a genuine confident model error.

Audit categories and decision rules. Upon review each of the 54 cases was assigned to exactly one of the following mutually exclusive categories, using the rules below *in order*:

1. **Grading or answer-key issue.** The model answer is semantically correct, or at least as defensible as the key, but is scored as incorrect because of a benchmark artifact. This includes exact-match formatting mismatches (e.g., `false` vs. `False`), option-letter vs. full-option-text mismatches, duplicate or semantically equivalent answer options, and incorrect answer keys. *Decision rule:* if the model answer is independently verifiable as correct or equivalent, assign this category.
2. **Ambiguous question or ambiguous label.** The prompt admits multiple defensible interpretations, or the benchmark label draws a boundary that is not uniquely determined by the problem statement. *Decision rule:* if a reasonable expert could defend both the model answer and the benchmark answer, assign this category.

3. **Genuinely wrong confident prediction.** The model answer is clearly inconsistent with the problem or gold option, and no grading artifact or ambiguity explains the discrepancy. *Decision rule:* assign this category only after (1) and (2) are ruled out. This is the residual category.

Headline audit result. Table 10 shows the pooled result. Of the 54 apparent HVR = 0 errors, 46 do not survive audit as genuine false accepts: 30 are grading or answer-key issues and 16 are ambiguous items. Only 8 cases remain as genuinely wrong confident predictions. Under this audit, the estimated false-accept rate of the HVR = 0 gate is therefore 8/1384 = 0.58%, rather than the raw 54/1384 = 3.90%.

Table 10. Audit of all 54 raw HVR = 0 disagreements.

Variant / category	Count	% of 54	% of N_0=1384
Raw disagreements	54	100.0%	3.90%
Grading or answer-key issue	30	55.6%	2.17%
Ambiguous question / label	16	29.6%	1.16%
Genuinely wrong confident	8	14.8%	0.58%
Raw precision	1330 / 1384	—	96.10%
Audit-adjusted precision	1376 / 1384	—	99.42%

The corresponding Wilson intervals are: raw precision $= 96.10\%$ (95% CI: [94.94%, 97.00%]) and audit-adjusted precision $= 99.42\%$ (95% CI: [98.86%, 99.71%]).

Breakdown by dataset and model. The 54 raw disagreements are not concentrated in a single benchmark. By dataset, the audit yields 23 BBH cases, 6 GPQA cases, and 25 MMLU-Pro cases. By final category, BBH contributes 7 grading/key issues, 12 ambiguous items, and 4 genuine errors; GPQA contributes 5 grading/key issues, 0 ambiguous items, and 1 genuine error; MMLU-Pro contributes 18 grading/key issues, 4 ambiguous items, and 3 genuine errors. By model, the corresponding counts are: Claude (3 / 4 / 1), Gemini (3 / 0 / 0), GPT-OSS-120B (4 / 3 / 0), GPT-OSS-20B (5 / 2 / 2), Grok (11 / 5 / 5), Qwen3 (3 / 0 / 0), and Qwen3-14B (1 / 2 / 0), where each triple is (grading/key, ambiguous, genuine).

Representative audited cases. The grading/key bucket is not dominated by borderline rescues; many are straightforward evaluation artifacts. Three representative examples are independently checkable.

First, in `mmlu_pro.gpt_oss_120b.11998`, the model selects option J (`+j50 ohms`) while the key marks option E (`-j50 ohms`). For a short-circuited transmission-line setting with $\beta d = \pi$, the standard relation $Z_{\text{in}} = jZ_0 \tan(\beta d)$ supports the model-side sign convention used in the worked solution rather than the provided key. Second, in `mmlu_pro.gpt_oss_120b.11904`, the model selects option C (*period*) while the key marks option I (*row number on periodic table*); these are semantically equivalent. Third, several cases are pure grading-format failures, including exact-match mismatches such as `false` vs. `False` on BBH Boolean Expressions and option-letter vs. full-option-text mismatches.

The ambiguous bucket mainly contains tasks whose label boundary is itself underspecified. A representative example is BBH Salient Translation Error Detection, where the distinction between a named-entity error and a factual error can be label-dependent even when the model's explanation is coherent.

The genuine-error bucket is small but real. The 8 surviving cases have mean verbalized confidence 0.978. They include three Dyck-language completions with an extra leading bracket, one clear output-format failure (`\boxed{54}` instead of an option letter), and four standard wrong-option selections with high stated confidence.

Interpretation. The main claim of the paper does not depend on the audit: the raw benchmark result is already 96.10% precision at 25.4% coverage. The audit sharpens the interpretation of the residual 3.90% raw error rate by showing that most of it comes from benchmark artifacts or intrinsically ambiguous labels rather than from genuinely wrong confident predictions. For this reason, we report both numbers in the paper: the raw benchmark precision as the primary result, and the audit-adjusted 99.42% figure as a secondary diagnostic estimate under the adjudication criteria above.

Scope, limitations, and sensitivity. This audit is a manual post-hoc analysis of the 54 raw disagreements only, conducted by a single annotator. It should therefore be read as a transparency measure on the failure modes of the HVR = 0 gate, not as

a replacement benchmark label set. We do not use the audit anywhere in model selection, thresholding, or score construction. Two structural safeguards partially mitigate subjectivity: the ordered decision rules ensure that category (3) is the residual after (1) and (2) are ruled out, so any classification ambiguity *inflates* the genuine-error count; and the 30 grading/key cases are independently verifiable against domain references without annotator judgment.

The main locus of subjectivity is the 16-item ambiguous bucket. To bound the effect of this judgment call, Table 11 reports three scenarios: the audit as annotated; a stricter setting in which all ambiguous cases are reclassified as genuine errors; and the zero-correction baseline using the raw labels only. Even under the stricter scenario, precision remains 98.27% (1360/1384), showing that the central claim does not hinge on trusting the annotator's judgment on every ambiguous item.

Table 11. Sensitivity of HVR = 0 precision to audit trust level.

Scenario	Genuine errors	Precision	Wilson 95% CI
As annotated	8	99.42%	[98.86%, 99.71%]
All ambiguous → genuine	24	98.27%	[97.32%, 98.89%]
Zero correction (raw labels only)	54	96.10%	[94.94%, 97.00%]

Appendix D: Sample Size Ablation

Performance plateaus near 60 traces for both models; we use 90 as the default to provide margin against sampling noise.

Table 12. Sample-size ablation under stratified sampling.

Traces	Sampling	Model	SD AUROC	HVR AUROC
15	stratified	Claude	— (empty markers)	—
30	stratified	Claude	0.821	0.671
60	stratified	Claude	0.829	0.751
90	stratified	Claude	0.833	0.756
120	stratified	Claude	0.831	0.748
180	stratified	Claude	0.827	0.743
15	stratified	Qwen3	0.779	0.783
30	stratified	Qwen3	0.776	0.784
60	stratified	Qwen3	0.776	0.783
90	stratified	Qwen3	0.774	0.782
120	stratified	Qwen3	0.775	0.783
180	stratified	Qwen3	0.774	0.784

Stratified sampling.

Table 13. Sample-size ablation under random pooled sampling.

Traces	Sampling	Model	SD AUROC	HVR AUROC
15	random_pooled	Claude	—	—
30	random_pooled	Claude	—	—
60	random_pooled	Claude	0.835	0.754
90	random_pooled	Claude	0.831	0.751
120	random_pooled	Claude	0.833	0.759
180	random_pooled	Claude	0.828	0.746
15	random_pooled	Qwen3	0.776	0.779
30	random_pooled	Qwen3	0.775	0.779
60	random_pooled	Qwen3	0.776	0.785
90	random_pooled	Qwen3	0.776	0.784
120	random_pooled	Qwen3	0.777	0.785
180	random_pooled	Qwen3	0.774	0.783

Random pooled sampling. Values are mean SD/HVR AUROC over BBH/GPQA/MMLU-Pro.

Appendix E: Embedding Model Ablation

Table 14. Embedding-model ablation for SELFDOUBT AUROC across datasets.

Embedding	Model	BBH	GPQA	MMLU	Mean
SELFDOUBT AUROC					
bge-m3	Claude	0.825	0.824	0.904	0.851
OpenAI-3-large	Claude	0.820	0.832	0.905	0.852
Qwen3-emb-8b	Claude	0.822	0.829	0.903	0.852
bge-m3	Qwen3	0.748	0.793	0.758	0.766
OpenAI-3-large	Qwen3	0.763	0.813	0.765	0.780
Qwen3-emb-8b	Qwen3	0.750	0.811	0.782	0.781

SELFDOUBT range: 0.001 (Claude), 0.015 (Qwen3). The maximum spread across embedding backends is 1.5 percentage points, confirming that marker discovery is near-immune to encoder choice. Given this robustness, we use `bge-m3` as the practical default because it is locally runnable, inexpensive, and empirically competitive.

Appendix F: Seed Set Size Ablation

Table 15. Seed-set size ablation on mean SELFDOUBT AUROC.

Seed Set	Mean SD AUROC	Δ vs top_10
top_10	**0.7895**	0.000
top_8	0.7861	−0.003
top_20	0.7826	−0.007
top_2	0.7822	−0.007
top_4	0.7815	−0.008
top_12	0.7814	−0.008
top_6	0.7813	−0.008
top_16	0.7803	−0.009
random_10	0.7782	−0.011

Performance is stable across coherence-ranked subsets (`top_2` through `top_20`), with a maximum spread of 0.8 pt. The gap to `random_10` (1.1 pt) confirms that coherence ranking contributes meaningful signal, but the method is not brittle to seed-set size within the ranked family. We select `top_10` as the default because it achieves the highest mean AUROC, while `top_8` performs comparably.

Appendix G: Cross-Dataset Marker Transfer (MuSR)

Per-model means at 90 MuSR (Sprague et al., 2024) traces:

Table 16. Cross-dataset marker transfer: per-model mean SELFDOUBT AUROC at MuSR@90.

Model	Original	MuSR@90	Δ
`gpt-oss-120b`	0.8406	0.8366	−0.004
`gpt-oss-20b`	0.7908	0.7867	−0.004
Qwen3	0.7745	0.7654	−0.009
Qwen3-14B	0.7876	0.7853	−0.002
Claude Sonnet 4.6	0.8358	0.8358	0.000
Grok 4.1 Fast	0.8101	0.8089	−0.001
Gemini 2.5 Flash	0.6871	0.6475	−0.040

Gemini's degradation is driven by BBH (−11.8 pt) while GPQA and MMLU-Pro are near-flat, consistent with its anomalous sparse-trace behavior. Excluding Gemini, mean degradation is 0.3 percentage points, confirming that the transferred markers capture model-specific language more than dataset-specific vocabulary.

Appendix H: Stage 2 vs. Seeds-Only

For HVR specifically, Stage 2 prevents large degradations on thought summary models:

Table 17. Stage 2 data-driven markers vs. seeds-only markers for HVR AUROC on summaries.

Model	Dataset	Data-Driven	Seeds Only	Δ (HVR)
Claude	BBH	0.733	0.659	-0.074
Claude	MMLU-Pro	0.834	0.700	$\mathbf{-0.134}$
Grok	MMLU-Pro	0.720	0.678	-0.042

SELFDOUBT AUROC shows a dead split (9–9 wins, +0.005 mean for seeds-only), indicating that z-score fusion absorbs moderate marker noise. The design implication is asymmetric: Stage 2 is mandatory for standalone HVR and the HVR = 0 gate, where marker precision directly controls false-accept risk, but optional for the fused SELFDOUBT score, where confidence fusion compensates.

Appendix I: Deployment Tier-2 Ranker Comparison

Table 18. Deployment Tier-2 ranker comparison on hard-subset ranking and cascade performance.

Tier-2 Ranker	Cascade AURAC	T2 AUROC (HVR>0)	Cost
z-sum (SD)	**0.8944**	**0.7572**	$O(1)$
HVR only	0.8883	0.6975	$O(1)$
SE	0.8830	0.7250	$O(N)$
Verb only	0.8756	0.7180	$O(1)$
TL+VB	0.8730	0.7149	$O(1)$
Random	0.8245	0.4917	—

Z-sum beats SE as Tier-2 ranker by 3.2 pt T2 AUROC and 1.1 pt cascade AURAC at $O(1)$ cost. HVR-only's cascade AURAC (0.8883) is propped up by the Tier-1 gate; its T2 AUROC on the hard subset (0.6975) is weakest among non-random rankers, validating the fusion design.

Appendix J: Subgroup Wilcoxon Tables

On full traces, SELFDOUBT significantly beats all three comparators on AUROC. On thought summaries, it loses to TL+VB (3–6 W–L) but remains significant against SE and Verb; this subgroup is where the aggregate TL+VB non-significance originates.

Table 19. Subgroup paired Wilcoxon results for SELFDOUBT vs. baselines on AUROC and AURAC.

Group	Comparison	n	Mean Δ AUROC	W–D–L	p
Full traces	vs TL+VB	12	+0.032	9–0–3	**0.026**
Full traces	vs SE	12	+0.042	10–0–2	**0.005**
Full traces	vs Verb	12	+0.051	9–1–2	**0.002**
Summaries	vs TL+VB	9	-0.010	3–0–6	0.787
Summaries	vs SE	9	+0.058	7–0–2	**0.037**
Summaries	vs Verb	9	+0.035	7–0–2	**0.049**
Full traces	vs TL+VB	12	+0.011 AURAC	8–0–4	**0.039**
Full traces	vs SE	12	+0.006 AURAC	9–0–3	0.102
Full traces	vs Verb	12	+0.029 AURAC	11–0–1	$<$**0.001**
Summaries	vs TL+VB	9	+0.032 AURAC	3–0–6	0.850
Summaries	vs SE	9	-0.007 AURAC	7–0–2	0.248
Summaries	vs Verb	9	+0.014 AURAC	7–0–2	**0.014**

Appendix K: Implementation and Hardware Details

Calibration pass. We use 90 traces per model, sampled randomly from the pooled BBH/GPQA/MMLU-Pro calibration set. Usable matched rows in the current artifacts are: Claude=90, `gpt-oss-20b`=86, `gpt-oss-120b`=90, Qwen3=80, Qwen3-14B=90, Grok=90, and Gemini=87.

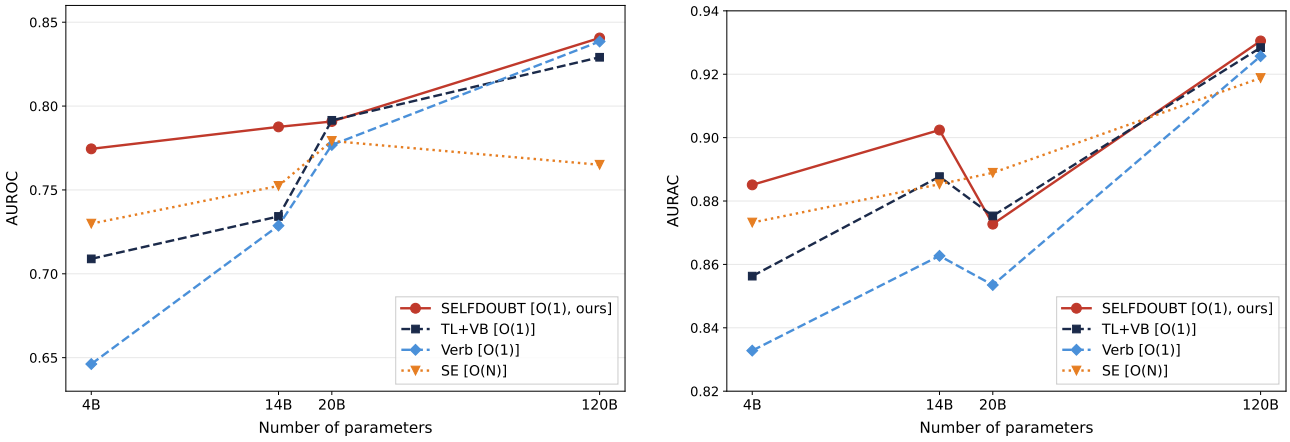
Inference. Marker matching uses pre-compiled word-boundary regex patterns (`\b...\b`, case-insensitive). HVR and z-sum computation require only arithmetic per query (4 stored scalars per model). Verbalized confidence is extracted from model output via a standard prompt suffix: "Confidence: [0-100]%" and then rescaled to $[0, 1]$.

Model serving. Qwen3-4B was run on a rented RTX 3090 GPU. All other closed-model evaluations were run through OpenRouter, which provides a unified API with multi-provider routing and automatic fallback across available providers (OpenRouter, c;b). OpenRouter exposes `logprobs` and `top_logprobs` as optional parameters, but parameter support is provider-dependent and must be checked at the model/provider level (OpenRouter, a). Because token-level log-probabilities were not available consistently for the OpenRouter-served model/provider combinations used in our runs, we do not report logit-based baselines.

Datasets. GPQA-Diamond (198 questions; Rein et al., 2023), BBH/Big-Bench Hard (300 stratified questions; Suzgun et al., 2022), MMLU-Pro (300 stratified questions; Wang et al., 2024).

Appendix L: AUROC and AURAC Scaling Plots

Figure 5 reports AUROC (left) and AURAC (right) as functions of model size on raw traces only. Across the four plotted methods, AUROC generally improves with scale, with the clearest monotonic trend coming from SELFDOUBT, which rises from 0.7745 (Qwen3-4B) to 0.8406 (GPT OSS 120B).



(a) AUROC **(b)** AURAC

Figure 5. Scaling vs. model size (raw traces only). **(a)** AUROC. **(b)** AURAC.

Relative to the baselines, SELFDOUBT remains competitive at every scale on both metrics. TL+VB and Verb generally improve with scale but start noticeably lower at small models, while SE is less monotonic across the same range. AURAC is somewhat noisier at intermediate scales, but SELFDOUBT still attains the strongest endpoint at 120B (0.9305).

Appendix M: Prompting and Extraction Details

M.1 Verbalized-confidence prompting. For verbalized-confidence runs, the effective prompt consists of the question text followed by a task-specific instruction block, separated by two newlines. Across all datasets, the confidence suffix is standardized as:

```
Confidence: [0-100]%
```

GPQA-Diamond.

```
After solving, put only the correct option in the box, and output the final answer
exactly once in the form \boxed{...}, followed by your confidence that the answer
is correct as a percentage.  Use this exact format:  Confidence:  [0-100]%
```

MMLU-Pro.

```
After solving, output your final answer exactly once as the option label in
parentheses, e.g., (C). Then output your confidence using this exact format:
Confidence:  [0-100]%
```

BBH task families. Rather than repeating 27 near-identical prompt strings, we group BBH tasks by their required answer format. In every case, the answer instruction below is followed by the same confidence suffix (`Confidence: [0-100]%`).

Table 20. BBH task-family answer-format instructions used in verbalized-confidence prompting.

Answer type	Tasks and instruction
Boxed `True`/`False`	`boolean_expressions`: output the final answer exactly once in the form `\boxed{...}` using only `True` or `False`.
Boxed `Yes`/`No`	`causal_judgement`, `navigate`, `web_of_lies`: output the final answer exactly once in the form `\boxed{...}` using only `Yes` or `No`.
Boxed lowercase `yes`/`no`	`sports_understanding`: output the final answer exactly once in the form `\boxed{...}` using only `yes` or `no`.
Boxed `valid`/`invalid`	`formal_fallacies`: output the final answer exactly once in the form `\boxed{...}` using only `valid` or `invalid`.
Boxed option label	`date_understanding`, `disambiguation_qa`, `geometric_shapes`, `hyperbaton`, `logical_deduction_three_objects`, `logical_deduction_five_objects`, `logical_deduction_seven_objects`, `movie_recommendation`, `penguins_in_a_table`, `reasoning_about_colored_objects`, `ruin_names`, `salient_translation_error_detection`, `snarks`, `temporal_sequences`, `tracking_shuffled_objects_three_objects`, `tracking_shuffled_objects_five_objects`, `tracking_shuffled_objects_seven_objects`: output the final answer exactly once in the form `\boxed{...}` using only the option label in parentheses, e.g., `(C)`.
Boxed integer	`multistep_arithmetic_two`: output the final answer exactly once in the form `\boxed{...}` using only the final integer (with sign if needed). `object_counting`: output the final answer exactly once in the form `\boxed{...}` using only the final integer.
Structured free-form output	`dyck_languages`: output the final answer exactly once as `Final Answer: <tokens>`, using only the missing bracket tokens separated by single spaces. `word_sorting`: output the final answer exactly once in the form `\boxed{...}`, using only the sorted words separated by single spaces.

M.2 Confidence parsing. Confidence is extracted from the answer region, defined as the text after the last `</think>` tag when such a tag is present, and otherwise from the full response. We first apply a strict parser that looks for an explicit `Confidence:` or `Confidence=` field followed by a numeric percentage. If no such field is found, we apply a fallback parser that scans the same answer region for percentage expressions more broadly. In both cases, we retain the last valid percentage in the range $[0, 100]$ and rescale it to $[0, 1]$.

M.3 Stage-1 seed-generation prompts. Stage 1 uses one prompt for hedging and one for verification.

Hedge prompt.

```
List 50 single English words that a speaker uses when they are uncertain or lack
confidence in a claim they are making.  Return only the words, one per line, no
numbering.
```

Verify prompt.

```
List 50 single English words that a speaker uses when they are performing a
concrete action to test or confirm an answer they already gave -- such as
recalculating a value, substituting it back into an equation, cross-checking
against a known fact, or computing an independent check.  These words should
indicate the speaker is actively testing correctness, not merely reconsidering or
reflecting.  Return only the words, one per line, no numbering.
```

M.4 Stage-1 generation models and voting rules. We query four models (`gpt-oss-20b`, `gpt-oss-120b`, `qwen3-14b`, `grok-4.1-fast`), requesting five independent runs per role.

Selection uses a strict-majority rule at two levels:

1. **Within model:** keep a word if it appears in at least 3 of 5 runs.
2. **Across models:** keep a word if it is retained by at least 3 of 4 models.

The resulting candidate sets are then passed through iterative coherence pruning using `BAAI/bge-m3` embeddings, cosine threshold 0.7, minimum keep count 10, and at most 6 pruning rounds. In the saved artifacts, the hedge role yields 41 global-majority candidates and retains all 41 after coherence filtering; the verify role yields 48 global-majority candidates, of which 46 survive after pruning.

M.5 Example retained seeds. Table 21 shows the top-10 retained words from the post-pruning coherence ranking for each role.

Table 21. Example retained Stage-1 seeds after coherence filtering.

Hedge	**Verify**
possibly	check
seemingly	reassess
maybe	reevaluate
apparently	re-evaluate
probably	reinspect
presumably	rechecking
perhaps	verifying
likely	reconfirming
reportedly	recheck
seems	prove

Appendix N: Threshold Sensitivity

Both Stage 2 classification thresholds (τ_{verify} and τ_{hedge}) are scaled jointly by a single multiplier applied to the per-model defaults in Appendix B. We sweep five multipliers $\{0.5\times, 0.75\times, 1.0\times, 1.25\times, 1.5\times\}$ and report mean SELFDOUBT AUROC across all 21 runs.

Table 22. Sensitivity of mean SELFDOUBT AUROC to joint Stage 2 threshold scaling.

Multiplier	τ_{verify} range	τ_{hedge} range	Mean SD AUROC	Δ vs. 1.0×
0.5×	[0.045, 0.085]	[0.025, 0.100]	0.7675	−0.022
0.75×	[0.068, 0.128]	[0.038, 0.150]	0.7807	−0.009
1.0× (default)	[0.090, 0.170]	[0.050, 0.200]	**0.7895**	0.000
1.25×	[0.113, 0.213]	[0.063, 0.250]	0.7854	−0.004
1.5×	[0.135, 0.255]	[0.075, 0.300]	0.7649	−0.025

Ranges reflect per-model variation from Appendix B; all models' thresholds are scaled by the same multiplier simultaneously. The total spread across all five conditions is 2.5 pt (0.7649–0.7895). Across the central three conditions (0.75×–1.25×), the spread narrows to 0.9 pt, confirming that SELFDOUBT is robust to threshold choice within a broad neighborhood of

the defaults. Degradation at the extremes ($0.5\times$ and $1.5\times$) is driven primarily by Qwen3, whose marker sets become either too inclusive or too restrictive when thresholds deviate by 50%. The z-score fusion absorbs moderate threshold perturbations: the SELFDOUBT AUROC range (2.5 pt) is substantially smaller than the HVR-alone range (12.6 pt) across the same conditions, confirming that the fusion design reduces sensitivity to the heuristic threshold selection.

Appendix O: Algorithmic Summary

This appendix gives compact pseudocode for the two implementation views used throughout the paper: SELFDOUBT calibration/scoring and the deployment-time accept/defer cascade.

Algorithm 1 SELFDOUBT: Calibration and Scoring

Calibration (once per model)
Require: 90 unlabeled traces $\{T_1, \ldots, T_{90}\}$ with verbalized confidences $\{V_1, \ldots, V_{90}\}$
Ensure: Dictionaries $\mathcal{H}, \mathcal{V}$; scalars $\mu_{\text{hvr}}, \sigma_{\text{hvr}}, \mu_{\text{v}}, \sigma_{\text{v}}$
Stage 1: Seed generation
Query 4 LLMs $\times$ 5 runs each for hedge/verify word lists
Retain words appearing in a majority of runs and a majority of models
Embed candidates with `bge-m3`; iteratively drop outliers below cosine 0.7 to the centroid (up to 6 rounds, floor of 10 words)
Select `top_10` ranked seeds $\rightarrow$ centroids $c_{\text{hedge}}, c_{\text{verify}}$
Stage 2: Per-model marker expansion
Extract 1–3-gram candidates from $\{T_i\}$; retain n-grams seen in at least min_trace_count traces
Classify by margin $\Delta(g) = \cos(g, c_{\text{verify}}) - \cos(g, c_{\text{hedge}})$ into $\mathcal{H}$ or $\mathcal{V}$
Z-score calibration
for each T_i **do**
  $h_i \leftarrow$ hedge occurrences in T_i matched against $\mathcal{H}$
  $v_i \leftarrow$ verify occurrences in T_i matched against $\mathcal{V}$
  $\text{hvr}_i \leftarrow h_i/(v_i + 1)$
end for
$\mu_{\text{hvr}}, \sigma_{\text{hvr}} \leftarrow \text{mean}, \text{std}(\{\text{hvr}_i\})$
$\mu_{\text{v}}, \sigma_{\text{v}} \leftarrow \text{mean}, \text{std}(\{V_i\})$

Scoring (per query)
Require: Trace T, verbalized confidence V
Ensure: SELFDOUBT score s
$h \leftarrow$ hedge occurrences in T matched against $\mathcal{H}$
$v \leftarrow$ verify occurrences in T matched against $\mathcal{V}$
$\text{hvr} \leftarrow h/(v + 1)$
$s \leftarrow \dfrac{-\text{hvr} - \mu_{\text{hvr}}}{\sigma_{\text{hvr}}} + \dfrac{V - \mu_{\text{v}}}{\sigma_{\text{v}}}$
return s

Algorithm 2 SelfDoubt Deployment Cascade (per query)

Require: Trace T, verbalized confidence V, dictionaries $\mathcal{H}, \mathcal{V}$, scalars $\mu_{\text{hvr}}, \sigma_{\text{hvr}}, \mu_{\text{v}}, \sigma_{\text{v}}$, threshold τ
Ensure: Decision $\in \{\text{ACCEPT}, \text{DEFER}\}$
$h \leftarrow$ number of word-boundary regex matches from $\mathcal{H}$ in T
$v \leftarrow$ number of word-boundary regex matches from $\mathcal{V}$ in T
if $h = 0$ **then** {Tier 1: HVR = 0 gate}
  return ACCEPT
end if
$\text{hvr} \leftarrow h/(v + 1)$ {Tier 2: scored ranking}
$s \leftarrow \dfrac{-\text{hvr} - \mu_{\text{hvr}}}{\sigma_{\text{hvr}}} + \dfrac{V - \mu_{\text{v}}}{\sigma_{\text{v}}}$
if $s \geq \tau$ **then**
  return ACCEPT
else
  return DEFER
end if