



SelfDoubt scores a single reasoning trace by combining the **Hedge-to-Verify Ratio**, a behavioral signal derived from hedging relative to self-verification, with the model's **verbalized confidence**. The two signals are standardized and added without training a predictor or using correctness labels. Across 21 model-dataset runs, SelfDoubt achieves the highest mean AUROC and AURAC, significantly outperforming Semantic Entropy on AUROC while statistically matching its AURAC with one generation instead of ten.

1 Background and research gap

Reasoning models are expensive to sample and often hide internals. That makes uncertainty quantification hard exactly in the setting where a single trace may already consume thousands of tokens.

What the trace adds

Reasoning traces expose revisions, doubts, and checks before the final answer. The key signal is not just whether doubt appears, but whether the model acts on that doubt before committing.

Family	Cost / access	Deployment limitation
SE / SV / GEO	O(10) samples	Repeated long generations
SEP / log-prob UQ	Hidden states or logits	Unavailable through many APIs
Verb / trace length	O(1), black-box	Inconsistent across models and tasks
SelfDoubt	O(1), trace text	No learned predictor or internals

Contributions

HVR. A trace-derived ratio of hedging to verification that captures the balance between expressing uncertainty and acting on it.	SelfDoubt. Training-free z-score fusion of HVR and verbalized confidence.
Unsupervised marker discovery. Model-specific hedge and verification dictionaries from 90 unlabeled traces.	Deployment cascade. A high-precision HVR=0 gate followed by thresholded continuous ranking.

2 Method: HVR, fusion, and marker discovery

Hedging expresses doubt: "maybe," "perhaps," "not sure." **Verification** acts on doubt: "let me check," "substitute back," "confirm." A hedge count alone cannot tell whether the model checked itself or merely accumulated uncertainty.

TRACE-DERIVED CHANNEL	REPORTED CHANNEL
$HVR(T) = \frac{h(T)}{(v(T) + 1)}$ Higher HVR means more hedging relative to verification, and is associated with a greater probability of error.	Confidence: 62% $C(T) = 0.62$ Verbalized confidence captures information that may not be visible in hedge and verification markers. Z-score normalization places both signals on comparable scales before fusion.

$$\text{SelfDoubt}(T) = z_{\text{cal}}(-\text{HVR}) + z_{\text{cal}}(C)$$

$z_{\text{cal}}(x) = (x - \mu_x) / \sigma_x$. The four calibration scalars are estimated once from the same **90 unlabeled traces/model**.

● ● ● one reasoning trace · no logprobs

“The mechanic greets the receptionist because she is standing in front of the door.” Who is *she*?
...refers to the receptionist... **hmm** **perhaps** the mechanic? ... **maybe** both are valid...
probably the receptionist but **maybe** ambiguous... **let me verify** the grammar rule... but
perhaps it is unclear...

Answer: receptionist · Confidence: 62%

h=7 · v=1 → HVR=7/(1+1)= **3.5** · C= **0.62**

Worked fusion:
 $z_H = z_{\text{cal}}(-\text{HVR})$, $z_C = z_{\text{cal}}(C)$
SelfDoubt = $z_H + z_C$

Build once, score every query

Once per model, we collect 90 unlabeled traces. The same sample is used to expand consensus seeds into model-specific hedge and verification dictionaries and to estimate μ and σ for oriented HVR and verbalized confidence. These frozen artifacts—two dictionaries and four calibration scalars—are reused at inference.

1 Single pass Obtain one trace, final answer, and reported confidence.	2 Scan trace Count model-specific hedge and verify markers.	3 Compute HVR Convert the two marker counts into $h/(v+1)$.	4 Parse confidence Map the final reported percentage to C in [0,1].	5 Calibrate + fuse Apply stored μ/σ values and sum both z-scores.
--	---	--	---	--

What the model-specific dictionaries learn

Static hedge lexicons neither capture model-specific phrasing nor distinguish expressions of uncertainty from concrete verification actions. Stage 2 expands consensus seeds into frequent model-specific 1–3-grams, then filters phrases that are not clearly closer to one behavioral centroid.

Model	Hedge examples	Verification examples
GPT-OSS 120B	perhaps · might be · possibly · probably · maybe · plausible	check · compute · examine · evaluate · determine · let's check
Qwen3 4B	perhaps · might be · hmm perhaps · probably · maybe · so perhaps	check · verify · me verify · recalculate · confirm

The dictionaries are frozen before evaluation and reused across BBH, GPQA, and MMLU-Pro.

3 Robustness, failure modes, and limitations

Why 90 traces? SelfDoubt performance changes little between 60 and 180 calibration traces for both Claude and Qwen3. Smaller samples occasionally produce empty marker sets, so we use 90 as a stable default.	Embedding sensitivity Across bge-m3, OpenAI, and Qwen embeddings, mean SelfDoubt varies by 0.001 on Claude and 0.015 on Qwen3.
Deployment calibration Replacing oracle test-set normalization with 90-trace calibration statistics costs only 0.0048 AURAC.	Cross-dataset transfer Markers discovered only from MuSR transfer to BBH, GPQA, and MMLU-Pro with a mean AUROC delta near −0.004 for 6 of 7 models.
Failure Modes On some Qwen3 runs, weak verbalized confidence dilutes the stronger HVR signal. For Gemini, compressed summaries expose few detectable hedge and verification markers, limiting HVR variation; its verbalized confidence is also weakly discriminative, with mean AUROC $\approx$ 0.590 across the three benchmarks.	

What the results support

Correctness tracks how a model expresses uncertainty and whether it engages in concrete verification, not simply how long it reasons. The pattern appears in full traces and survives provider-compressed summaries, although compression weakens continuous discrimination.

Limitations. HVR coverage depends on model and provider style; suppressed hedging or superficial verification can distort the score. Evaluation is currently limited to English multiple-choice reasoning. When verbalized confidence is unavailable, SelfDoubt reduces to HVR alone.

0.7895

best mean AUROC across all methods

96.1%

accuracy when HVR = 0 at 25.4% coverage

71%

coverage at 89.7% accuracy in deployment

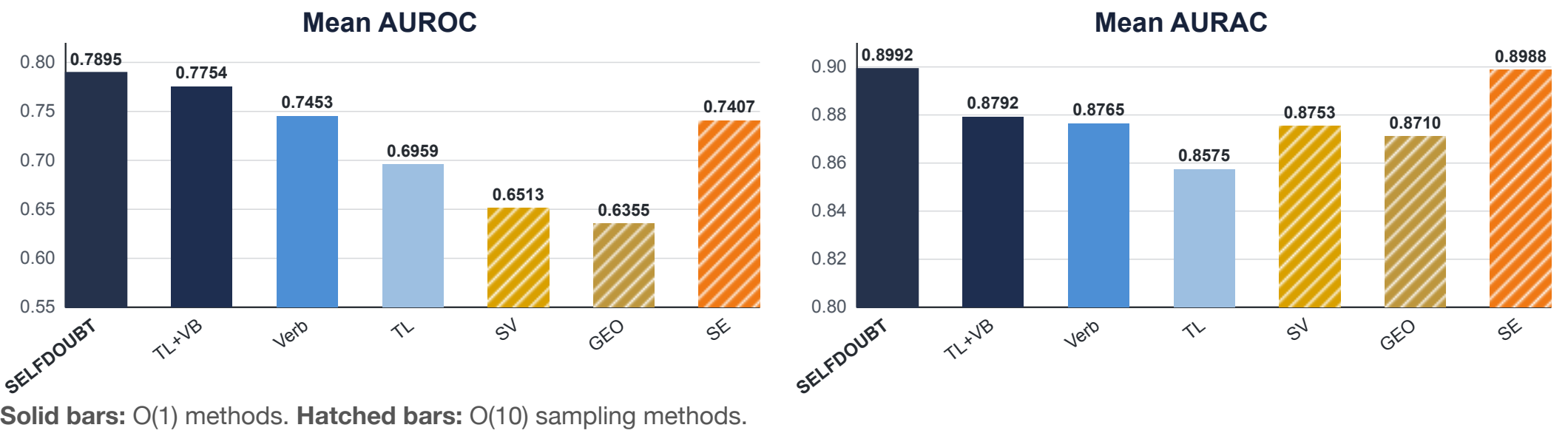
4 Experimental setup and trace access

We evaluate **7 reasoning models × 3 multiple-choice reasoning benchmarks = 21 runs**. GPQA-Diamond tests graduate-level science, BBH covers diverse symbolic and logical tasks, and MMLU-Pro uses 10-option professional and academic questions. All methods are evaluated on matched samples; sampling baselines use 10 generations per query.

Raw reasoning traces · 12 runs Qwen3 4B, Qwen3 14B, GPT-OSS 20B, GPT-OSS 120B. Local inference exposes the full generated reasoning text, including revisions and self-checks.	Thought summaries · 9 runs Claude Sonnet 4.6, Grok 4.1 Fast, Gemini 2.5 Flash. Proprietary APIs expose compressed thought summaries rather than all reasoning tokens.
---	--

Trace access	SelfDoubt AUROC	SelfDoubt AURAC	HVR=0 coverage	HVR=0 accuracy
Raw traces	0.7984	0.8977	17.7%	96.0%
Thought summaries	0.7777	0.9013	35.4%	96.2%

5 Comparison across single-pass and sampling-based methods



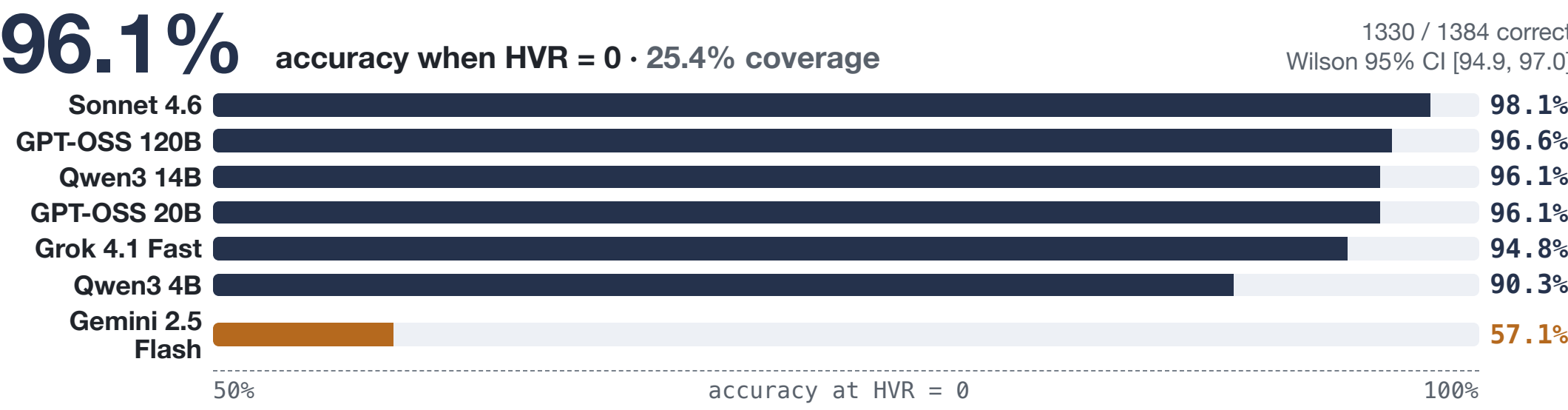
Solid bars: O(1) methods. Hatched bars: O(10) sampling methods.

SelfDoubt obtains the highest mean AUROC and AURAC across the 21 matched runs: 0.7895 AUROC and 0.8992 AURAC. The AUROC improvement over Semantic Entropy is significant (+0.049, paired Wilcoxon p=0.001, 17–4 wins); the AURAC difference is not (0.8992 vs 0.8988), meaning SelfDoubt matches its selective-prediction quality at substantially lower inference cost. HVR alone reaches 0.7509 AUROC and 0.8883 AURAC. By trace access, SelfDoubt AUROC is higher on raw traces than thought summaries (0.7984 vs. 0.7777), while AURAC is nearly unchanged (0.8977 vs. 0.9013).

Cross-trace consistency is the practical result. SelfDoubt is strongest on raw traces and remains competitive on compressed summaries; SE keeps high AURAC on summaries but loses discrimination. Against TL+VB, aggregate differences are not significant, but SelfDoubt wins significantly on the raw-trace subgroup.

6 The zero-hedge regime

What HVR = 0 means HVR=0 means no hedge markers were detected in the trace. Because v+1 is always positive, HVR equals zero exactly when h=0 — not proof that the model experienced no doubt internally. Empirically, this zero-hedge subset is correct 96.1% of the time at 25.4% coverage.	Precision first The gate accepts only traces with no detected hedge markers; the continuous SelfDoubt score ranks all remaining cases.	Coverage is the tradeoff Zero-hedge coverage depends on model and trace format. Compressed summaries may omit uncertainty markers, so coverage rises or falls with provider style.
--	--	--

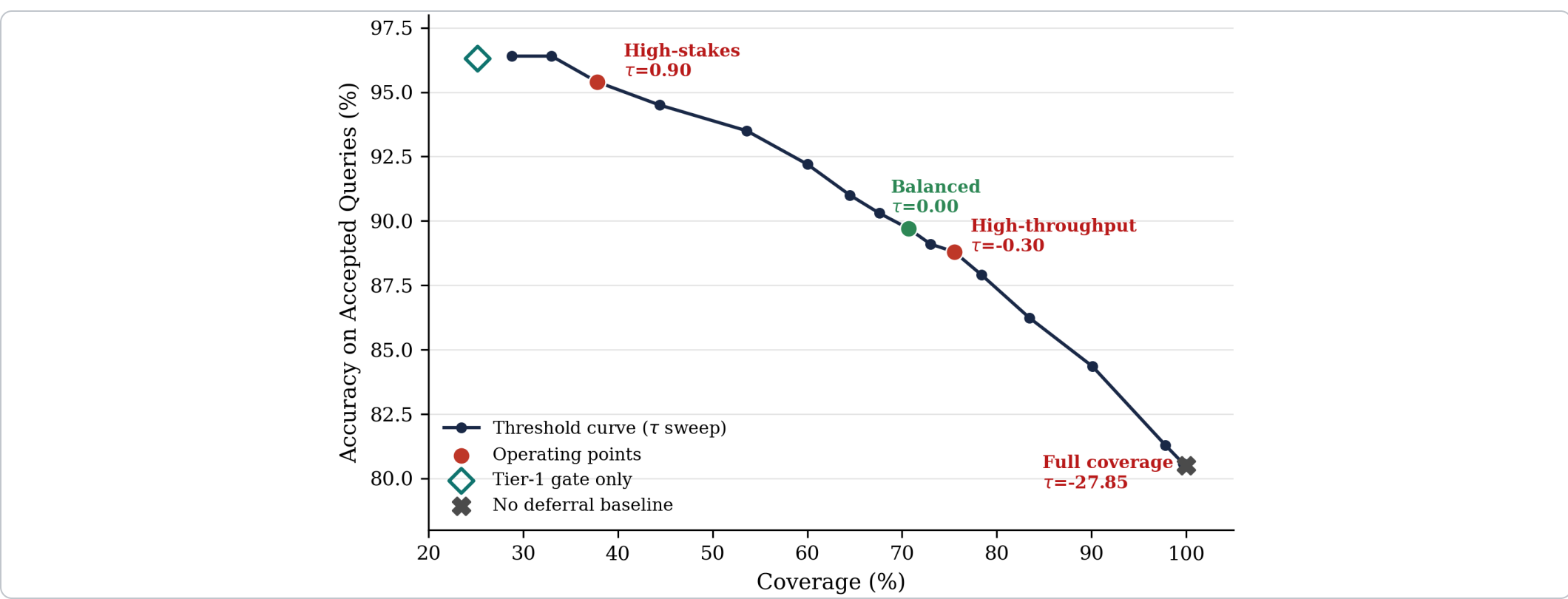


Six of seven models exceed 90%. Gemini has only seven zero-hedge examples and is the sparse-summary outlier. Zero-hedge precision is stable across raw traces and thought summaries (96.0% vs. 96.2%), but coverage is strongly provider-dependent, ranging from 53.3% for Claude to 0.9% for Gemini.

Audit of 54 raw disagreements: 30 grading/key issues · 16 ambiguous labels · **8 genuine errors** (0.58% of the gate).

A post-hoc audit yields a label-noise-corrected precision of 99.4%. We retain 96.1% as the primary benchmark result and treat the audit as supporting evidence.

7 Deployment: an accept-defer cascade



Tier 1 · structural gate If HVR=0 and at least four zero-hedge examples occurred during calibration, accept immediately. Under this guard, Tier 1 answers 25.2% of queries at 96.3% accuracy.	Tier 2 · calibrated ranker For HVR>0, accept when $z_{\text{cal}}(-\text{HVR}) + z_{\text{cal}}(C) \geq \tau$; otherwise defer. At $\tau=0$, coverage is 70.7% at 89.7% accuracy.
---	---

The no-deferral baseline is 80.5% accuracy. The cascade therefore adds 9.2 percentage points at the balanced operating point and degrades smoothly as coverage increases.

Representative operating points

Operators choose τ according to the cost of false acceptance versus deferral. Higher thresholds preserve accuracy; lower thresholds recover throughput. The curve is monotonic, so this choice does not depend on a single hand-picked point.

Policy	Coverage	Accuracy on accepted
Tier 1 only	25.2%	96.3%
$\tau = 0.90$	37.8%	95.4%
$\tau = 0.60$	53.6%	93.5%
$\tau = 0.30$	64.5%	91.0%
$\tau = 0.00$ · balanced	70.7%	89.7%