Filtered Reasoning Score: Evaluating Reasoning Quality on a Model's Most Confident Traces

Manas Pathak, Xingyao Chen, Shuozhe Li, Amy Zhang, Liu Leqi
University of Texas at Austin
{manaspathak, caspar.cxy06, shuozhe.li, leqiliu}@utexas.edu

Abstract

Should we trust Large Language Models (LLMs) with high accuracy? LLMs achieve high accuracy on reasoning benchmarks, but correctness alone does not reveal the quality of the reasoning used to produce it. This highlights a fundamental limitation of outcome-based evaluation: models may arrive at correct answers through flawed reasoning, and models with substantially different reasoning capabilities can nevertheless exhibit similar benchmark accuracy, for example due to memorization or over-optimization. In this paper, we ask: given existing benchmarks, can we move beyond outcome-based evaluation to assess the quality of reasoning itself? We seek metrics that (1) differentiate models with similar accuracy and (2) are robust to variations in input prompts and generation configurations. To this end, we propose a reasoning score that evaluates reasoning traces along dimensions such as faithfulness, coherence, utility, and factuality. A remaining question is how to aggregate this score across multiple sampled traces. Naively averaging them is undesirable, particularly in long-horizon settings, where the number of possible trajectories grows rapidly, and low-confidence correct traces are more likely to be coincidental. To address this, we introduce the Filtered Reasoning Score (FRS), which computes reasoning quality using only the top-K% most confident traces. Evaluating with FRS, models that are indistinguishable under standard accuracy exhibit significant differences in reasoning quality. Moreover, models with higher FRS on one benchmark tend to perform better on other reasoning benchmarks, in both accuracy and reasoning quality. Together, these findings suggest that FRS complements accuracy by capturing a model's transferable reasoning capabilities. We open source our evaluation codebase: https://github.com/HumainLab/filtered_reasoning_score_evaluation.

Two correct answers, very different reasoning quality

Trace A *Score: 100/100*

Problem: Octagon with same perimeter as hexagon of side 16 cm. Find octagon side length.

Perimeter $= 6 \times 16 = 96$ cm. Octagon has 8 sides: $8s = 96 \Rightarrow s = 12$. $\boxed{12}$

Trace B *Score: 25/100*

Problem: GCF of 6432 and 132, increased by 11.

Lists factors, concludes GCF = 4, gets 4+11=15 *[wrong]* → ... $\boxed{23}$

Figure 1: Two traces from different models produce correct final answers and receive the same Pass@1 score, yet their reasoning quality is vastly different.

1 Introduction

Large language models have advanced rapidly in recent years, with newer models achieving higher scores on an expanding set of reasoning benchmarks (Hendrycks et al., 2021; Cobbe

et al., 2021; Rein et al., 2024). Yet the way we evaluate these models has not kept pace. The dominant paradigm remains final-answer accuracy: a model is scored by how often it produces the correct output, with no regard for the reasoning process that produced it.

This paradigm is increasingly inadequate. Models can produce flawed reasoning that still leads to correct answers (Lightman et al., 2024; Uesato et al., 2022; Turpin et al., 2023). As a result, accuracy gains do not reliably reflect improvements in reasoning quality (Xia et al., 2025), and benchmark saturation further reduces their ability to distinguish between models (Deveci et al., 2025). Moreover, outcome-based evaluation can be sensitive to prompt choice and generation configuration, further obscuring differences in underlying reasoning ability (Hochlehnert et al., 2025).

These limitations motivate a shift from evaluating outcomes to evaluating reasoning itself. In this work, we ask: *can we use existing benchmarks to assess the quality of reasoning, beyond final-answer correctness?* We seek metrics that (1) differentiate models with similar accuracy with respect to their underlying reasoning capabilities, and (2) are robust to variations in prompts and generation configurations. To this end, we propose to evaluate reasoning traces directly using a rubric-based reasoning score that measures properties such as faithfulness, coherence, utility, and factuality (Section 3.1).

A second challenge is aggregation: models can produce many reasoning traces for the same problem, with substantial variation in both likelihood and quality. While averaging over all sampled traces may provide an unbiased estimate of its expected reasoning score, this is not the quantity most relevant in practice. Deployed systems typically act on one selected output, often chosen explicitly or implicitly based on confidence (Kang et al., 2025; Wen et al., 2025; Geifman and El-Yaniv, 2017). Evaluating all traces equally therefore gives disproportionate influence to low-probability trajectories that are unlikely to be selected in practice.

To address this, we introduce the **Filtered Reasoning Score (FRS)**, which evaluates reasoning quality on the most confident subset of traces in terms of their faithfulness, coherence, utility, and factuality. For each problem, we sample multiple reasoning traces, estimate per-trace confidence from token-level probabilities, and retain only the top-K% by confidence before computing the final score. This design focuses evaluation on the high-confidence region of the model's reasoning trace distribution, where outputs are more representative of the model's intended inference behavior. A high FRS therefore requires both strong reasoning and high confidence on high-quality reasoning traces (Section 3.3).

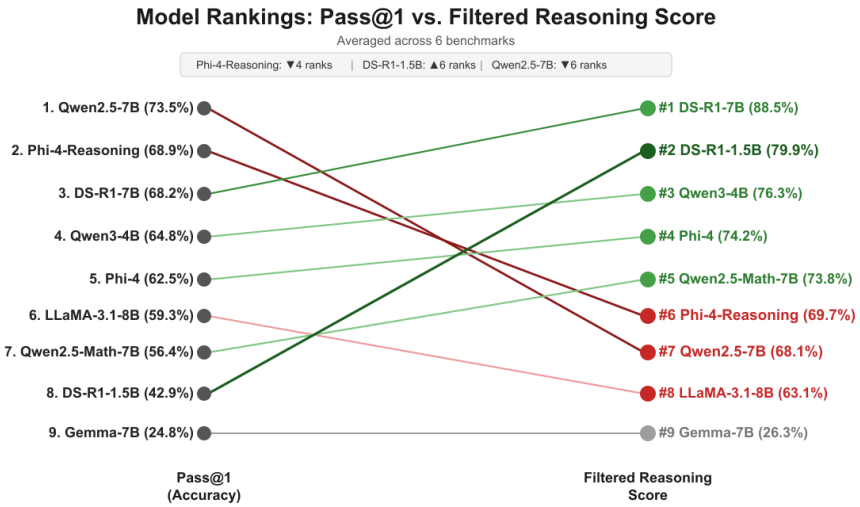

Figure 2: Model rankings under greedy pass@1 (left) vs. Filtered Reasoning Score with $K = 10\%$ (right), averaged across 6 benchmarks. Green lines indicate models that rise under FRS; red lines indicate models that fall. Qwen2.5-7B: #1 $\rightarrow$ #7; DS-R1-1.5B: #8 $\rightarrow$ #2.

Empirically, FRS complements accuracy by identifying distinguishing characteristics among models with similar accuracy. We evaluate FRS across 9 open-weight models (1.5B to 14B parameters) and 6 reasoning benchmarks (Figure 2). On MATH500 (Lightman et al., 2024), two models with identical greedy accuracy (63.6%) differ by 16.5 FRS points. The highest-accuracy model overall drops from #1 to #7 under FRS, while a model ranked #8 by accuracy rises to #2. Across benchmarks, models within 0.7% accuracy differ by up to 18.8 FRS points. These differences reflect systematic variation in how models assign confidence over reasoning traces: some models confidently produce flawed reasoning that may lead to correct answers, while others more reliably assign high confidence to well-reasoned solutions. Such distinctions are invisible under accuracy-based evaluation, yet directly impact which reasoning is generated in practice.

Our contributions are:

1. **We identify confidence-conditioned reasoning quality as a distinct evaluation target beyond answer accuracy.** Correctness alone does not capture reasoning quality: two traces can reach the same answer while differing substantially in faithfulness, coherence, utility, and factuality. This matters especially when LLM systems prioritize high-confidence outputs, making it important to evaluate whether a model's most confident reasoning traces are of high quality.
2. **We introduce the Filtered Reasoning Score (FRS).** FRS evaluates reasoning quality on the most confident traces, requiring both strong reasoning and high confidence on well-reasoned solutions.
3. **We show that FRS reveals structure hidden by accuracy-based evaluation.** Across 9 open-weight models and 6 reasoning benchmarks, FRS exposes ranking reversals, large separations among accuracy-similar models, and systematic differences in how confidence relates to reasoning quality, with all key results supported by non-overlapping bootstrap 95% CIs. FRS is also the only metric among six candidates that significantly predicts whether confidence-based selection improves or degrades reasoning quality (r=0.49, p<0.001; Appendix U).

2 Related Work

Reasoning evaluation and benchmark saturation. Most reasoning benchmarks evaluate final-answer correctness (Cobbe et al., 2021; Hendrycks et al., 2021; Rein et al., 2024), but accuracy improvements do not reliably track reasoning quality (Xia et al., 2025; Mondorf and Plank, 2024). Process-level evaluation has received growing attention through process reward models (PRMs) (Lightman et al., 2024; Uesato et al., 2022) and rubric-based taxonomies (Lee and Hockenmaier, 2025); (Turpin et al., 2023) show that chain-of-thought explanations (Wei et al., 2022) can be unfaithful to the model's actual reasoning (Li et al., 2026a;b). PRMs are the closest existing alternative to our rubric-based scorer; we compare against a PRM head-to-head in Appendix V and find that the two are uncorrelated over the full trace distribution ($\rho = 0.15$) but converge once traces are confidence-filtered ($\rho = 0.78$), indicating that the confidence filter, not the scorer, is the source of FRS's predictive signal. Meanwhile, benchmark saturation is eroding the discriminative power of existing evaluations (Deveci et al., 2025; Hashemi Amroabadi et al., 2025). Our approach addresses saturation differently: rather than creating harder problems, we condition evaluation on a model's own confidence signal, exposing structure invisible to any accuracy threshold.

Confidence estimation and calibration. LLM confidence estimation spans verbalized methods (Xiong et al., 2024; Tian et al., 2023), sampling-based approaches such as self-consistency (Wang et al., 2022) and semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), and logit-based methods (Hendrycks and Gimpel, 2017; Kadavath et al., 2022); Geng et al. (2024) provide a comprehensive survey. Recent work on targeted token aggregation (Fu et al., 2025b;a; Devic et al., 2025) informs our confidence estimator design (Section 3.2). On calibration, Zeng et al. (2025) find that RL-trained models achieve better confidence calibration, and Xie et al. (2026) show that post-RL SFT can restore calibration degraded by GRPO training. FRS is complementary: calibration asks whether confidence tracks *correctness*; FRS asks whether high confidence selects for strong *reasoning*, a question calibration

cannot answer, since models can reach correct answers through flawed reasoning. Selective prediction (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017; 2019; Wen et al., 2025) studies when to abstain but does not evaluate the quality of accepted traces; FRS fills this gap.

LLM-as-a-judge. We use GPT-4o-mini as a rubric-based judge following the paradigm established by Zheng et al. (2023) and extended by Liu et al. (2023). We validate against GPT-4o, Claude Sonnet 4.5, and human annotators (Appendix D).

3 Methodology

3.1 Reasoning Quality Evaluation

To capture the quality of a model's reasoning process, we score chain-of-thought (CoT) traces along four dimensions following the taxonomy of Lee and Hockenmaier (2025): **faithfulness** (internal consistency without hidden shortcuts), **coherence** (logical flow between steps), **utility** (whether each step contributes to the solution with correct calculations), and **factuality** (grounding in the problem context without hallucinations). We provide the full scoring rubric in Appendix A. Empirically, the four dimensions organize into two axes: faithfulness, coherence, and utility are mutually correlated (Spearman $\rho \in [0.90, 0.98]$) and capture a dominant logical-quality axis, while factuality is largely independent ($\rho \in [0.35, 0.61]$ with the other three) and contributes a distinct grounding signal. A leave-one-dimension-out ablation shows FRS is stable to removing any single logical-axis dimension, and we retain all four following Lee and Hockenmaier (2025); see Appendix B for the full analysis.

We use GPT-4o-mini as an automated evaluator following the LLM-as-a-judge paradigm (Zheng et al., 2023). The judge receives the question, the model's full response (CoT trace and final answer), the ground-truth answer, and the scoring rubric, and produces a score from 1 (very poor) to 5 (excellent) for each dimension. We validate judge reliability in three ways: cross-model agreement with GPT-4o (82% within ± 1) and Claude Sonnet 4.5 (75% within ± 1), and a blind human annotation study with 10 annotators and 500 labeled samples, achieving 78% within ± 1 agreement and a Spearman correlation of $\rho = 0.73$ (Appendix D). The **Reasoning Score** is the average across the four dimensions, normalized to $[0, 1]$:

$$\text{ReasoningScore}(r) = \frac{1}{4} \sum_{d \in \{\text{faith., coh., util., fact.}\}} \frac{s_d(r) - 1}{4} \tag{1}$$

where $s_d(r) \in \{1, 2, 3, 4, 5\}$ is the judge's score for dimension d on reasoning trace r.

3.2 Per-Trace Confidence Estimation

The first stage of FRS is to assign each reasoning trace a scalar confidence score that can be computed without correctness labels. We use a logit-based estimator derived from token-level probabilities. Rather than averaging over the full trace, we focus on its low-probability tail, since these tokens concentrate the model's uncertainty, whereas many high-probability tokens are generic and contribute little to distinguishing strong traces from weak ones.

Concretely, for each input x_i, the model generates a reasoning trace $r_i = (t_1, \ldots, t_{L_i})$. For each token, we extract its conditional probability $P(t_j) = P(t_j \mid x_i, t_1, \ldots, t_{j-1})$. We then define trace-level confidence as

$$C(r_i) = \frac{1}{|\mathcal{T}_{\text{low}}|} \sum_{t \in \mathcal{T}_{\text{low}}} P(t), \tag{2}$$

where $\mathcal{T}_{\text{low}}$ is the set of tokens in r_i with probabilities below the pth percentile.

We use p=10% as the default high-confidence operating point, targeting the region where low-probability tokens most concentrate the model's uncertainty. This value is chosen once via a coarse calibration sweep and is stable across nearby choices in the 5–15% range

(Appendix C). We note that this estimator is empirically rather than theoretically motivated, and none of the paper's conclusions are load-bearing on it: rankings are qualitatively consistent across $p \in \{10, 20, 30, 40, 50\}$ and across alternative confidence estimators, including full-trace mean log-probability ($\rho = 0.983$) and self-consistency confidence computed purely from answer agreement without any logit access ($\rho = 0.80$; Appendices C, T). These confidence scores define the ranking used in the next stage of FRS, where we retain only the most-confident traces for reasoning evaluation.

3.3 Filtered Reasoning Score

The second stage of FRS uses the confidence scores above to construct a filtered set of traces, and then measures reasoning quality only on that set. For each problem x_i, we sample k=16 reasoning traces at temperature T=0.7 using chain-of-thought prompting, compute $C(r_i^{(j)})$ for each trace, and pool all traces within a model–benchmark pair. We then rank these traces by confidence and retain the top K% as the *filtered set*.

We use K=10 as the default, targeting the region where a model's outputs would be trusted most in deployment; rankings are qualitatively consistent across $K \in \{10, 20, 30, 40, 50\}$ (Appendix G). Because traces are pooled globally before ranking, one might worry that the filtered set concentrates on easy problems where the model is uniformly confident. We verify this is not the case: more than 64% of problems contribute at least one trace to the top-10% set, and the set draws *more* from hard problems (22.3% of selected traces come from problems solved at most 4/16 times) than from uniformly easy ones (16.4% from problems solved 16/16); see Appendix H.

To estimate FRS efficiently, we partition the top 50% of traces into five disjoint confidence bins (0–10%, 10–20%, . . . , 40–50%) and randomly sample 50 traces per bin for each model–benchmark pair. These sampled traces are scored by GPT-4o-mini using the rubric in Section 3.1. The Filtered Reasoning Score at threshold K is then

$$\mathrm{FRS}_K = \frac{1}{|\mathcal{S}_K|} \sum_{r \in \mathcal{S}_K} \mathrm{ReasoningScore}(r), \tag{3}$$

where $\mathcal{S}_K$ denotes all sampled traces from bins within the top K%.

This construction makes the two components of FRS explicit: it rewards models that produce strong reasoning, but only when that reasoning appears in the traces that their own confidence ranks highest. A model with strong reasoning but misaligned confidence or weak reasoning overall will therefore score poorly.

4 Experiments

4.1 Setup

We evaluate 9 open-weight models spanning 1.5B to 14B parameters: DeepSeek-R1-Distill-Qwen-1.5B and -7B (Guo et al., 2025), LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct and Qwen2.5-Math-7B (Yang et al., 2024), Gemma-7B (Gemma Team et al., 2024), Phi-4 and Phi-4-Reasoning (Abdin et al., 2024), and Qwen3-4B in thinking mode (Qwen Team, 2025), covering general-purpose, math-specialized, and reasoning-optimized architectures. We evaluate on GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021; Lightman et al., 2024), SVAMP (Patel et al., 2021), AQuA (Ling et al., 2017), GPQA (Rein et al., 2024), and CommonsenseQA (Talmor et al., 2019), spanning arithmetic, mathematical, commonsense, and graduate-level science reasoning. For each problem, we sample k=16 reasoning traces at T=0.7; confidence and FRS are computed as described in Sections 3.2–3.3.

4.2 Why Reasoning Quality, Not Just Accuracy

Before conditioning on confidence, we establish that reasoning quality adds information beyond accuracy. We show that the reasoning score is more stable and converges faster than accuracy.

Reasoning score produces more stable rankings. We measure three stability metrics (average pairwise Spearman ρ, percentage of rank inversions, and average maximum rank swing) across four evaluation conditions (direct prompting, few-shot, CoT with T=0, CoT with T=0.7). On 5 of 6 benchmarks, reasoning score produces higher rank correlations, fewer inversions, and smaller rank swings than accuracy. Averaged across benchmarks, reasoning-based rankings achieve Spearman ρ of 0.72 vs. 0.49 for accuracy, with 33% fewer inversions (19.9% vs. 29.9%) and 31% smaller maximum rank swings (2.4 vs. 3.5 positions). Full results are in Table 3 (Appendix F).

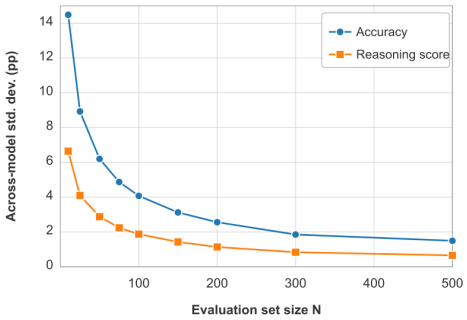

Figure 3: Median across-model std. dev. vs. evaluation set size N (54 model-dataset combinations). The reasoning score has lower variance for all sample sizes. Per-benchmark breakdowns in Appendix E.

Reasoning score converges faster. Reasoning scores reach stable rankings with fewer evaluation problems than accuracy (Figure 3). Across 54 model-dataset combinations, the median standard deviation of reasoning scores is lower than that of accuracy at every sample size, and the gap is largest at small N: at N=25, accuracy fluctuates at roughly 9pp while reasoning score is already below 4pp. However, unfiltered reasoning score is insufficient: models with similar averages can have opposite confidence-quality relationships. For example, two models may both average 85 on reasoning quality, yet one improves to 97 in its high-confidence region while the other drops to 65. Exposing this structure requires conditioning on confidence, which is what FRS does.

4.3 Why Confidence-Based Filtering Is Necessary

Having established that reasoning quality is a richer signal than accuracy, we address a key aggregation challenge: how should reasoning quality be summarized across multiple sampled traces? Averaging equally over all traces is not the quantity most relevant in practice. Deployed systems typically act on one selected output, often chosen explicitly or implicitly based on confidence (Kang et al., 2025; Wen et al., 2025; Geifman and El-Yaniv, 2017), giving disproportionate influence to low-probability trajectories unlikely to be surfaced in deployment. We therefore ask: *is better reasoning concentrated in the model's high-confidence region?*

Figure 4 plots average reasoning quality at each threshold $K \in \{50, 40, 30, 20, 10\}$ for three representative models. For DS-R1-7B, restricting to higher-confidence traces yields markedly better reasoning: its score rises from 85.7 at K=50% to 88.5 at K=10%. DS-R1-1.5B shows an even steeper gain, climbing from 72.2 to 79.9. By contrast, Phi-4-Reasoning moves in the opposite direction, dropping from 81.5 to 69.7, meaning its most-confident traces contain *worse* reasoning than its average output.

Two models with the same average reasoning score can therefore differ sharply in the quality their confidence selects. Unfiltered averaging cannot distinguish a model whose confidence surfaces its best reasoning from one that surfaces its worst. We evaluate reasoning quality only in the high-confidence region, where deployed systems actually operate. The next section fixes K=10% and reports FRS across all models and benchmarks.

5 FRS Results: What Confidence Actually Selects

We now evaluate reasoning quality at K=10% across all 9 models and 6 benchmarks. Table 1 reports the full results, and Figure 2 shows the resulting ranking changes.

FRS is most informative where accuracy is least discriminative. Among the 34 per-benchmark model pairs with greedy accuracy gaps of at most 5 pp (out of 216 total), FRS

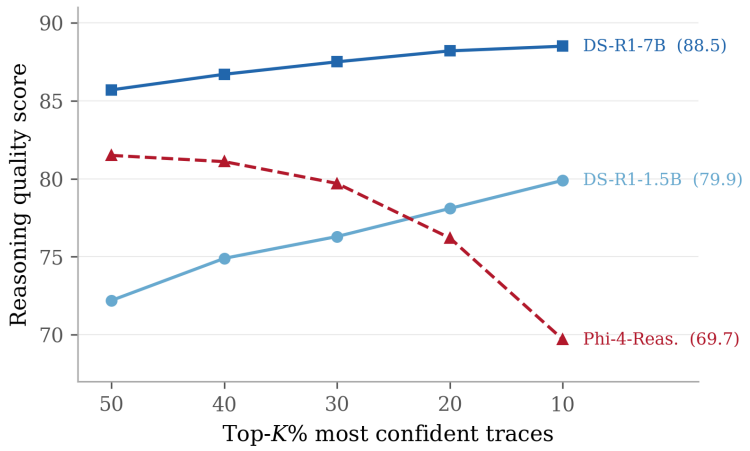

Figure 4: Reasoning quality on the top K% most-confident traces, averaged across 6 benchmarks, for three representative models. As the confidence filter tightens (lower K), quality improves for some models and degrades for others. Full results for all 9 models are in Appendix G.

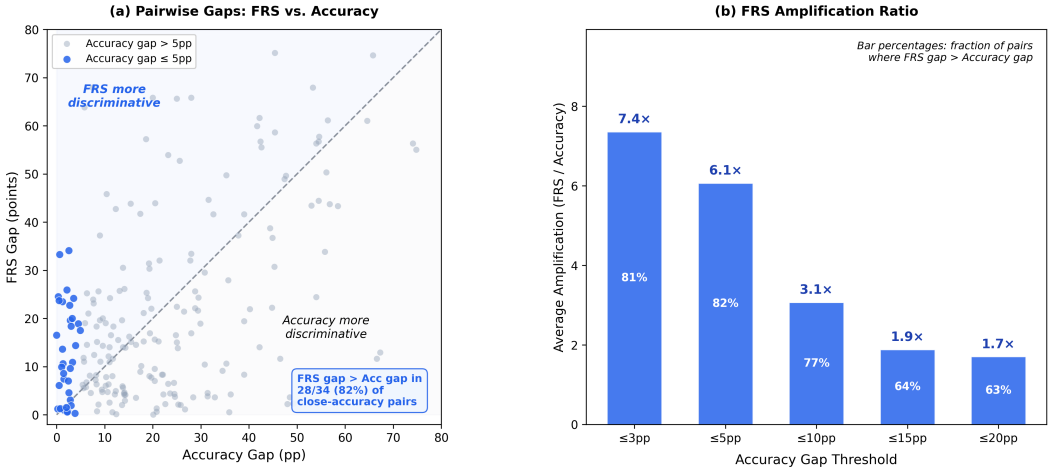

Figure 5: (a) Pairwise gaps: FRS vs. accuracy for all 216 per-benchmark model pairs. Blue points denote the 34 pairs with accuracy gap $\leq$5 pp; FRS produces a larger gap in 82% (28 of 34) of these close-accuracy cases. (b) FRS amplification ratio (mean FRS gap / mean accuracy gap) at increasing accuracy-gap thresholds. Bar percentages show the fraction of pairs where FRS gap exceeds accuracy gap. FRS is most discriminative where accuracy is least informative.

produces a larger separation in 82% of cases (28 of 34), with a mean gap roughly 6$\times$ greater than the corresponding accuracy gap. Among pairs within 3 pp, the amplification rises to 7.4$\times$ (Figure 5). Concrete examples confirm this: on MATH, DS-R1-7B and Qwen2.5-Math are exactly tied at 63.6% greedy accuracy yet differ by 16.5 FRS points (Figure 10, Appendix J). Similarly, at the aggregate level DS-R1-7B and Phi-4-Reasoning are nearly tied on pass@1 yet differ by 18.8 FRS points.

Top-confidence accuracy is still not enough. A natural objection is to simply measure accuracy on the most-confident traces. This does not suffice. Phi-4-Reasoning achieves the highest top-10% accuracy (79.3%), slightly above DS-R1-7B (77.1%), yet their FRS scores differ sharply: 69.7 vs. 88.5. The reason is that Phi-4-Reasoning's confident traces often contain correct answers wrapped in degenerate reasoning: short solution cores followed

Table 1: FRS at 10% filtering (reasoning quality, 0–100) across 6 benchmarks. Acc Avg is greedy decoding pass@1 accuracy. Models are sorted by average FRS. $\pm$ values denote bootstrap standard deviation of the mean (10,000 resamples over 50 scored traces per bin); sensitivity to the upstream sampling budget is analyzed in Appendix R.

Model	GSM8K	MATH	SVAMP	AQuA	GPQA	CSQA	FRS Avg	Acc Avg
DS-R1-7B	98.9$\pm$1.1	94.0$\pm$2.0	99.0$\pm$0.7	96.8$\pm$1.3	62.9$\pm$2.5	79.2$\pm$3.3	**88.5**$\pm$0.8	68.2
DS-R1-1.5B	95.1$\pm$2.2	84.5$\pm$3.6	96.9$\pm$1.7	90.1$\pm$2.5	56.6$\pm$4.1	56.1$\pm$4.2	**79.9**$\pm$1.3	42.9
Qwen3-4B	90.9$\pm$2.3	63.5$\pm$2.7	91.0$\pm$1.9	78.5$\pm$3.2	58.8$\pm$2.2	75.0$\pm$2.5	**76.3**$\pm$1.0	64.8
Phi-4	91.5$\pm$2.7	74.4$\pm$4.3	75.5$\pm$4.2	78.5$\pm$3.5	48.4$\pm$3.4	77.2$\pm$3.8	**74.2**$\pm$1.5	62.5
Qwen2.5-Math	90.8$\pm$2.5	77.5$\pm$4.1	74.5$\pm$4.4	74.9$\pm$4.1	54.6$\pm$3.0	70.6$\pm$3.4	**73.8**$\pm$1.5	56.4
Phi-4-Reas.	84.5$\pm$2.6	80.0$\pm$3.5	74.8$\pm$3.4	83.2$\pm$3.0	62.3$\pm$4.9	33.4$\pm$5.1	**69.7**$\pm$1.6	68.9
Qwen2.5-7B	65.6$\pm$2.2	75.6$\pm$2.6	64.9$\pm$1.9	77.2$\pm$2.4	55.1$\pm$2.7	70.1$\pm$3.5	**68.1**$\pm$1.1	73.5
LLaMA-3.1-8B	79.9$\pm$3.4	60.6$\pm$4.3	80.0$\pm$3.2	52.9$\pm$3.9	50.5$\pm$2.7	55.0$\pm$3.0	**63.1**$\pm$1.4	59.3
Gemma-7B	41.2$\pm$5.1	18.9$\pm$2.0	31.1$\pm$4.6	22.2$\pm$2.1	30.9$\pm$1.7	13.4$\pm$3.2	**26.3**$\pm$1.4	24.8

by repetitive continuations that inflate token-level confidence while degrading reasoning quality (Appendix M).

FRS reshuffles accuracy-based rankings. The per-benchmark effects above compound into substantial ranking shifts at the model level. Qwen2.5-7B, the highest-accuracy model overall at 73.5%, drops from #1 to #7 under FRS, indicating that its confidence does not reliably prioritize its best reasoning. DS-R1-1.5B moves in the opposite direction, rising from #8 by accuracy to #2 under FRS despite having the second-lowest accuracy (42.9%), because it consistently aligns high confidence with strong reasoning. All key ranking reversals involve FRS gaps exceeding 8 points with non-overlapping bootstrap 95% confidence intervals.

These conclusions are robust across filtering thresholds, confidence estimators, judge models, and sampling budgets. Rankings are nearly unchanged under full-trace mean log-probability ($\rho \geq 0.983$) and remain strongly correlated under self-consistency, a logit-free estimator ($\rho = 0.80$; Appendix T). Full ablations are in Appendices D, C, and R.

FRS predicts whether confidence-based selection helps or hurts. Beyond producing different rankings, FRS is the only metric that predicts a deployment-relevant outcome. For 54 model–benchmark pairs, we compare the reasoning quality of a model's top-confidence trace against a random baseline (Appendix U). FRS significantly predicts this *selection gain* (r=0.49, p<0.001); pass@1, high-confidence accuracy, and unfiltered reasoning score are all nonsignificant ($p > 0.14$). The result is not an artifact of judge coupling: rescoring the selection-gain outcome with a held-out, different-family judge (Claude Haiku 4.5) preserves the effect (r=0.367, 95% CI $[0.13, 0.56]$; Appendix U.5), and it is not reducible to response style, since FRS remains predictive after jointly controlling for trace length, lexical diversity, and 4-gram repetition (partial r=0.308, p=0.023; Appendix U.6).

Conditioning on confidence is what produces the signal. Confidence-agnostic evaluation of reasoning quality is not sufficient. We compare FRS against two confidence-agnostic baselines under the same judge: a single random trace per problem and the unfiltered average over all traces (Appendix U.4). Both amplify close-accuracy gaps at roughly 4$\times$ and neither predicts selection gain (r=0.01 and r=0.008, both n.s.), whereas FRS amplifies at 7.4$\times$ and predicts selection gain at r=0.49; bootstrap intervals over the $\leq$3 pp amplification are disjoint ($[7.04, 8.82]$ vs. $[3.46, 5.64]$). Evaluating only the single most-confident trace, the $K \to 1$ limit of confidence-conditioned evaluation, largely agrees with FRS (Spearman $\rho = 0.75$, amplification 7.9$\times$), with FRS providing a lower-variance estimate of the same quantity. This confirms that FRS captures a property that no confidence-agnostic metric measures: whether confidence-based selection actually surfaces better reasoning.

FRS correlates across benchmarks. Models with higher FRS on one benchmark tend to have higher accuracy and reasoning quality on other benchmarks. Figure 6 reports pairwise Spearman correlations $\rho(\mathrm{FRS}_i, \cdot_j)$ across models for $i \neq j$, with grand off-diagonal means

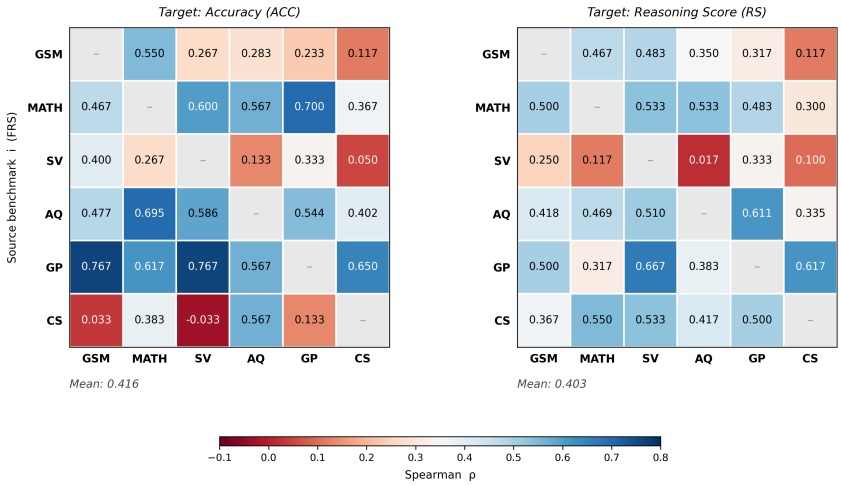

Figure 6: Spearman correlations $\rho(\mathrm{FRS}_i, \cdot_j)$ across models for $i \neq j$. Rows are source benchmark i (FRS), columns are target benchmark j. FRS has a mean $\rho = 0.416$ and $\rho = 0.403$ with Pass@1 (Accuracy) and Reasoning Score, respectively.

of $\rho = 0.416$ (vs. accuracy) and $\rho = 0.403$ (vs. reasoning score). GPQA and AQuA yield the strongest row-averages (up to $\rho = 0.674$), while SVAMP and CSQA produce weaker correlations (as low as $\rho = 0.163$), suggesting that the strength of cross-benchmark transfer varies by domain.

In addition, we conduct a leave-one-benchmark-out (LOBO) analysis: for each held-out benchmark, we average each model's FRS over the remaining five benchmarks and correlate the result with the held-out FRS. Because DS-R1-7B and DS-R1-1.5B share a distillation training process and may produce correlated FRS across benchmarks, we report a family-controlled variant, holding out both DS-R1 models jointly, as the primary number: the mean Spearman correlation is $\rho = 0.495$ across the remaining 7 models, positive in all six folds. Without the family control, the mean correlation over all 9 models is $\rho = 0.712$ (std. 0.158), with all six folds positive and five of six significant by permutation test ($p < 0.05$). The strongest generalization appears for GSM8K ($\rho = 0.933$, $p < 0.001$) and AQuA ($\rho = 0.820$, $p = 0.004$); the weakest is CSQA ($\rho = 0.467$, $p = 0.102$), consistent with commonsense reasoning placing different demands on confidence–quality alignment. These results indicate partial, not perfect, transfer: confidence–quality alignment carries cross-benchmark structure that is invisible to accuracy, and knowing a model's FRS on five benchmarks is informative about its FRS on the sixth.

6 Discussion

FRS as a practical audit for confidence-based selection. Many deployment strategies (e.g., abstention and test-time compute scaling) implicitly assume that a model's highest-confidence outputs are also its best ones. Our results show this does not hold uniformly. FRS provides a simple pre-deployment audit: if FRS improves under tighter filtering, confidence-based selection will likely surface strong reasoning; if it degrades, the same strategy may surface weaker reasoning at precisely the moments the model appears most certain. The cost is modest: rankings at k=8 traces per problem remain nearly identical to those at k=16 (Appendix R). As shown in Section 5, this holds across confidence estimators, suggesting confidence–quality alignment is a model property rather than an artifact of aggregation.

FRS fills a gap in the evaluation stack. Accuracy measures what a model gets right, but not whether the reasoning behind it is sound. Correct answers can arise from flawed or degenerate traces, so outcome-based metrics alone cannot assess reasoning quality in the high-confidence region where deployed systems operate. As our selection gain analysis confirms

(Section 5), this gap is deployment-relevant: FRS is the only metric among six candidates that significantly predicts whether confidence-based selection improves or degrades reasoning quality, a result that replicates under a held-out different-family judge (Appendix U.5) and diverges from what process reward models capture over the full trace distribution (Appendix V). Moreover, our leave-one-benchmark-out analysis shows that confidence–quality alignment carries stable cross-benchmark structure (family-controlled $\rho = 0.495$ across held-out folds, positive in all folds; $\rho = 0.712$ without the family control), not a benchmark-specific artifact. Together, these results establish confidence-conditioned reasoning quality as a distinct evaluation target for reasoning models in confidence-mediated settings.

Scope and limitations. FRS targets the broad class of confidence-mediated selection: best-of-N by self-certainty, confidence-based abstention, and confidence-thresholded test-time scaling. It does not assume that high-confidence traces are always the deployment-relevant outputs; rather, that premise is exactly what FRS audits, and Appendix K shows models where the confidence–correctness relationship is inverted. Inference pipelines built on verifier reranking, debate, or search change only the selection rule, not the rubric: FRS can be applied to whichever trace such a pipeline selects, and self-consistency, itself a confidence signal, is already covered (Appendix T.1). Our evaluation spans arithmetic, competition math, algebraic word problems, graduate-level science, and commonsense reasoning; demonstrating confidence–quality alignment directly on agentic planning, interactive tool use, code reasoning, and long-horizon decision-making remains an important direction, and the long-horizon case is the regime our motivation most directly targets, where low-confidence correct traces are most likely coincidental. Finally, the low-probability confidence estimator can be inflated by degenerate repetition (Appendix M); we treat this as a limitation of the confidence signal that the rubric-based judge corrects, and note that FRS is an evaluation metric rather than a reward signal, so exploiting this artifact would require optimization pressure the evaluation setting lacks.

What drives confidence–quality alignment? One striking pattern is that models whose FRS improves under tighter filtering all used reinforcement learning during post-training, whereas models whose FRS degrades did not. We view this as suggestive rather than conclusive: it is consistent with prior findings that RL improves confidence calibration, but our experiments were not designed to isolate RL from architecture, scale, or training data. We treat confidence–quality alignment as an important phenomenon revealed by FRS, and its mechanisms as an open question. If this alignment can be trained directly, FRS may serve not only as an evaluation metric but as a training target for improving reasoning models.

7 Conclusion

We introduced the Filtered Reasoning Score (FRS), a metric that evaluates reasoning quality on a model's most-confident outputs. Conditioning evaluation on confidence reveals structure that accuracy cannot: models indistinguishable under accuracy, and even under top-confidence accuracy, can differ sharply in the reasoning their confidence selects. FRS is most discriminative exactly where accuracy is least informative.

Evaluation should move beyond correctness when models are deployed in confidence-mediated settings. Accuracy measures whether a model can produce the right answer, but not whether its most-confident outputs reflect sound reasoning. FRS targets this question directly, and our leave-one-benchmark-out analysis confirms that the reasoning property it measures, confidence–quality alignment, partially transfers across benchmarks even under family-controlled evaluation.

An important direction for future work is understanding and improving confidence–quality alignment. Our results show this alignment varies systematically across models and training regimes, raising the possibility of treating confidence-conditioned reasoning quality not only as an evaluation target but as a training objective, encouraging models to assign high confidence selectively to well-reasoned outputs.

Ethics Statement

This work introduces the **Filtered Reasoning Score (FRS)**, an evaluation metric designed to assess *confidence-conditioned reasoning quality*. We view FRS as an auditing tool for modern deployment settings where systems act on a model's most-confident outputs. In such settings, correctness alone is insufficient: models can produce correct answers through flawed or degenerate reasoning, and confidence can preferentially surface these failures. By explicitly measuring reasoning quality under confidence-based selection, FRS aims to reduce misplaced trust in high-confidence outputs and improve the reliability of model evaluation.

The primary positive impact of FRS is to expose failure modes that are invisible to accuracy and even to top-confidence accuracy. By identifying whether a model's confidence aligns with strong reasoning, FRS enables practitioners to make more informed decisions about model selection, filtering strategies, and deployment policies. We therefore position FRS as a practical pre-deployment diagnostic: when FRS is high, confidence-based selection is more likely to surface sound reasoning; when FRS is low, such strategies may amplify poor reasoning despite high apparent confidence.

At the same time, FRS is not a complete measure of reliability or safety. A model with high FRS may still be unsafe, biased, or poorly calibrated in real-world settings. In addition, because FRS relies on LLM-based judges, it may inherit biases from the judge model and favor reasoning styles that are more legible to the evaluator. Like any metric, FRS may also be subject to optimization or gaming. We therefore emphasize that FRS should be used as *one component* of a broader evaluation stack, alongside accuracy, calibration, robustness testing, and human oversight, rather than as a standalone deployment criterion.

We take steps to mitigate these risks. We validate FRS across judge models and human annotations, report robustness across filtering thresholds and sampling budgets, and open-source the FRS evaluation pipeline to promote transparency and reproducibility. We also encourage users to evaluate FRS across diverse benchmarks and domains, as confidence–quality alignment may vary across settings.

FRS also raises fairness considerations. Benchmarks, prompts, and judge models may encode narrow distributions of tasks or linguistic norms, which can affect FRS scores across models or domains. As a result, FRS-based rankings should not be used in isolation for socially consequential decisions (e.g., in healthcare, education, or legal contexts). Instead, FRS should serve as a diagnostic signal within a broader, human-centered evaluation process.

Our experiments primarily use public benchmarks and model-generated outputs, and do not require collection of personal user data. For the human annotation study used to validate FRS (Appendix D.2), all 10 annotators were informed of the study's purpose and how their annotations would be used, and provided consent before participating; annotation was conducted blind to the automated judge's scores, no personal or sensitive data was collected from annotators, and annotators participated voluntarily. We also respect dataset and model licenses and aim to ensure reproducibility through release of the FRS codebase.

Finally, we note that FRS requires multi-sample generation and evaluation, which incurs additional computational cost compared to accuracy. In practice this cost is modest: the full main evaluation across all 9 models and 6 benchmarks requires 13,500 judged traces at a total judging cost of approximately $12 with GPT-4o-mini, and FRS requires no generation beyond standard best-of-N pipelines, since it operates on traces already produced during inference (Appendix W). We therefore report efficiency trade-offs and show that FRS rankings remain stable under reduced sampling. We view this cost as justified by the improved visibility FRS provides into how confidence interacts with reasoning quality. Overall, FRS contributes toward more responsible evaluation by shifting attention from whether models are correct to whether they are reliable *when they appear most confident*.

References

Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

İbrahim Ethem Deveci et al. The ouroboros of benchmarking: Reasoning evaluation in an era of saturation. *arXiv preprint arXiv:2511.01365*, 2025.

Siddharth Devic et al. Trace length is a simple uncertainty signal in reasoning models. *arXiv preprint arXiv:2510.10409*, 2025.

Abhimanyu Dubey et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Deep think with confidence: Reasoning LLMs with adaptive confidence-based early stopping. *arXiv preprint arXiv:2505.08895*, 2025a.

Yichao Fu, Junda Chen, Yonghao Zhuang, Zheyu Fu, Ion Stoica, and Hao Zhang. Reasoning without self-doubt: More efficient chain-of-thought through certainty probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025b. URL https://openreview.net/forum?id=wpK4IMJfdX.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2151–2159, 2019.

Gemma Team et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Kober, Horst Bischof, and Besmira Nushi. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Sayed Masoud Hashemi Amroabadi et al. Revitalizing saturated benchmarks: A weighted metric approach for differentiating large language model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *NeurIPS*, 2021.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1789–1814, Suzhou, China, 2025. Association for Computational Linguistics.

Shuozhe Li, Jincheng Cao, Bodun Hu, Aryan Mokhtari, Leqi Liu, and Amy Zhang. Care-rft: Confidence-anchored reinforcement finetuning for reliable reasoning in large language models. *arXiv preprint arXiv:2602.00085*, 2026a.

Shuozhe Li, Vaishnav Tadiparthi, Kwonjoon Lee, Nakul Agarwal, Hossein Nourkhiz Mahjoub, Ehsan Moradi Pari, Lizhang Chen, Amy Zhang, and Leqi Liu. Learning robust reasoning through guided adversarial self-play. *arXiv preprint arXiv:2602.00173*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, 2023.

Philipp Mondorf and Barbara Plank. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *NAACL*, 2021.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting world knowledge. In *NAACL*, 2019.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Bingbing Wen, Yunshan Cao, Yun Xiao, Zhenya Liu, and Philip S. Yu. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 2025.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730, 2025.

Xiaohu Xie, Xiaohu Liu, and Benjamin Yao. Know when you're wrong: Aligning confidence with correctness for LLM error detection. *arXiv preprint arXiv:2603.06604*, 2026.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. Thinking out loud: Do reasoning models know when they're right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1407. Association for Computational Linguistics, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

A Scoring Rubric for Reasoning Quality

We provide the full prompt given to the GPT-4o-mini judge for scoring reasoning traces.

REASONING QUALITY SCORING PROMPT

You are an expert evaluator of mathematical and logical reasoning. Score the chain-of-thought (CoT) on 4 dimensions using the scoring criteria below.
Each score must be an integer from 1–5 (1 = very poor, 5 = excellent).

1. FAITHFULNESS (1–5)
Definition: Reasoning is internally consistent, follows logical rules, and stays focused on the problem without hidden shortcuts or leaps.
- **5:** Perfect logical consistency, no contradictions, stays completely on-topic
- **4:** Minor inconsistencies or slight tangents, but overall coherent
- **3:** Some logical gaps or moderate off-topic content
- **2:** Significant logical flaws or frequent tangents
- **1:** Major contradictions, illogical leaps, or completely off-topic

Dock points for: Contradictory statements, logical leaps without justification, going off-topic, hidden assumptions, unjustified final answers, shortcut reasoning.

2. UTILITY (1–5)
Definition: Each step meaningfully contributes to solving the problem, calculations are correct, and reasoning efficiently leads to the final answer.
- **5:** Every step is necessary and correct, efficient path to solution
- **4:** Most steps useful, minor inefficiencies or small errors
- **3:** Some useful steps mixed with unnecessary ones or calculation errors
- **2:** Many unnecessary steps or significant calculation errors
- **1:** Mostly useless steps, major calculation errors, or repetitive content

Dock points for: Incorrect calculations, repetitive statements, unnecessary verbose explanations, steps that don't advance toward the solution, circular logic.

3. COHERENCE (1–5)
Definition: Steps flow smoothly from one to the next with clear logical progression and smooth transitions.
- **5:** Perfect flow, each step naturally follows from the previous
- **4:** Good flow with minor awkward transitions
- **3:** Some disjointed steps but overall progression
- **2:** Choppy flow with unclear connections between steps
- **1:** Disjointed, random steps with no clear progression

Dock points for: Abrupt transitions, missing connecting logic, disjointed sequence, poor organization, dangling references, disordered reasoning chain.

4. FACTUALITY (1–5)
Definition: Every step must be factually correct and grounded in the problem context, not hallucinated from surface-level understanding.
- **5:** All facts and statements are accurate and grounded in the problem
- **4:** Mostly accurate with minor factual errors
- **3:** Some factual errors or unsupported claims
- **2:** Multiple factual errors or significant hallucinations
- **1:** Major factual errors, hallucinations, or completely unsupported claims

Dock points for: Hallucinated facts, incorrect interpretations, unsupported assumptions, surface-level understanding leading to wrong facts, claims contradicting problem evidence.

EVALUATION PROCESS
1. Read the problem carefully to understand the context and given information
2. Analyze each step of the CoT reasoning
3. Check each step against the four criteria above
4. Assign scores based on the specific guidelines for each dimension

5. Ensure every step is evaluated for factual accuracy and logical soundness

Input fields: {problem}, {cot}, {gold}, {flags_summary}, {evidence}

Required output (JSON):
{"faithfulness": <1-5>, "utility": <1-5>, "coherence": <1-5>, "factuality": <1-5>}

B Rubric Dimension Analysis

A natural question is whether the four rubric dimensions capture distinct information or are functionally redundant. We analyze the dimension structure across all 54 model–benchmark pairs.

Pairwise structure. The four dimensions organize into two axes. Faithfulness, coherence, and utility are strongly mutually correlated (Spearman $\rho \in [0.90, 0.98]$), forming a dominant logical-quality axis. Factuality is largely independent of this axis, correlating at $\rho \in [0.35, 0.61]$ with each of the other three dimensions, and contributes a distinct grounding signal. Consistent with this structure, the model-level FRS ranking induced by faithfulness alone correlates with the full four-dimension ranking at $\rho = 0.994$, with coherence and utility behaving similarly ($\rho \geq 0.99$), while factuality alone yields $\rho = 0.861$ and produces meaningful ranking shifts when included.

Why factuality differs from faithfulness in practice. The two dimensions penalize different failure modes by construction (Appendix A). Faithfulness penalizes internal inconsistency: contradictions, unjustified leaps, and hidden shortcuts, regardless of whether the trace's premises are grounded in the problem. Factuality penalizes hallucinated premises, misread problem context, and unsupported claims, even when the subsequent chain is internally consistent. On the largely arithmetic benchmarks in our suite these failure modes frequently co-occur, since an error that breaks grounding usually also breaks internal consistency, which is why the logical-quality axis dominates. The dimensions nevertheless dissociate in both directions: on knowledge-heavy benchmarks such as GPQA, a trace can be internally coherent while resting on a hallucinated fact, and the Phi-4-Reasoning traces in Appendix M exhibit the converse pattern directly, scoring 5 on factuality while faithfulness and utility collapse to 2 or below, because every claim is grounded in the problem even though the reasoning degenerates into repetition.

Leave-one-dimension-out ablation. We recompute the paper's two headline quantities, close-accuracy amplification (Section 5) and LOBO transfer, under each three-dimension rubric subset.

Table 2: FRS behavior under leave-one-dimension-out rubric subsets. Amplification is the mean FRS gap over the mean accuracy gap among model pairs within the stated accuracy threshold; Transfer is the mean LOBO Spearman ρ.

Rubric Subset	Amp. $\leq$3 pp	Amp. $\leq$5 pp	Transfer LOBO ρ
Full (4 dimensions)	7.4×	6.1×	0.71
Drop faithfulness	7.5×	6.2×	0.71
Drop coherence	6.6×	5.5×	0.62
Drop utility	6.5×	5.4×	0.63
Drop factuality	9.5×	7.9×	0.75

FRS is stable to removing any single logical-axis dimension, with amplification in the 6.5–7.5× range and transfer in the 0.62–0.71 range, so no individual dimension is load-bearing. Dropping factuality *raises* amplification to 9.5× because grounding errors are rare on these largely arithmetic benchmarks, so removing the factuality signal sharpens separation along the logical-quality axis. We nevertheless retain all four dimensions following the taxonomy

of Lee and Hockenmaier (2025): factuality is the one axis carrying independent information ($\rho = 0.861$ vs. ≥ 0.99 for the others), and it is the relevant signal for knowledge-heavy benchmarks and deployment auditing, where hallucinated-but-coherent reasoning is precisely the failure mode outcome metrics miss.

C Low-Probability Token Analysis

What are low-probability tokens? The most frequent low-probability tokens (Figure 7, left) fall into two categories:

1. **Decision-point tokens**: Words that are largely interchangeable without affecting reasoning logic, such as "Okay," "Alright," and "Just." Low probability at these tokens reflects a choice among multiple equally valid continuations, indicating a natural branch point in the generation process. Low probability at these tokens is frequent and normal across models.
2. **Uncertainty-expressing tokens**: Words that explicitly signal confusion or doubt, such as "messed," "confuse," "misunderstood," and "Sometimes." Low probability at these tokens reflects genuine uncertainty in the reasoning process.

Both categories represent points where the model's reasoning is under stress, either because multiple paths are available or because the model is uncertain about how to proceed.

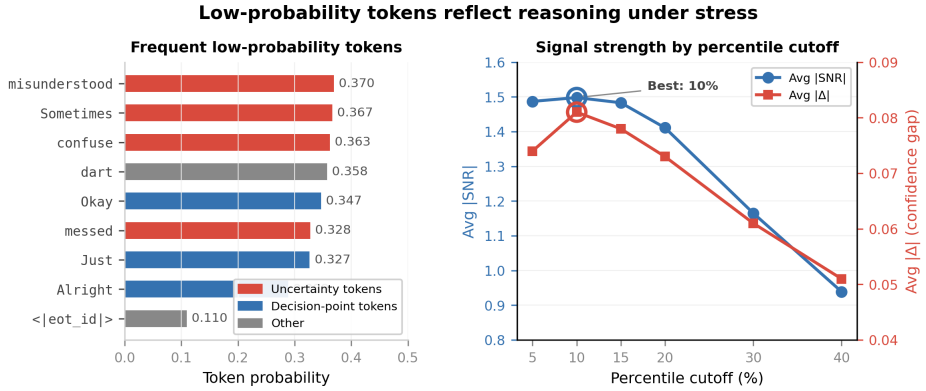

Figure 7: Left: Frequent low-probability tokens, categorized as uncertainty-expressing (red) or decision-point (blue). Right: Confidence signal quality (SNR and mean confidence gap) as a function of percentile cutoff. The 10% cutoff yields the strongest signal on both measures.

Analyzing the percentile cutoff. We evaluate percentile cutoffs from 5% to 40% using two separation metrics across all 9 models and 6 benchmarks. The first is the signal-to-noise ratio (SNR), defined as:

$$\mathrm{SNR} = \frac{\mathbb{E}[C \mid Y{=}1] - \mathbb{E}[C \mid Y{=}0]}{\sqrt{\frac{1}{2}\left(\mathrm{Var}[C \mid Y{=}1] + \mathrm{Var}[C \mid Y{=}0]\right)}} \tag{4}$$

with $Y_i \in \{0, 1\}$ denoting correctness. We also compute the average absolute difference in mean confidence between correct and incorrect populations (Figure 7, right).

The lowest 5–15% of tokens yield the strongest separation, with the 10% cutoff achieving the highest values on both measures (Avg $|\mathrm{SNR}| = 1.498$, Avg $|\Delta| = 0.081$). Beyond 15%, both metrics decline monotonically as high-probability tokens dilute the signal. The 5–15% range is consistently strong; we use 10% as the default throughout the paper because it provides the best overall signal while including enough tokens per trace for a stable estimate.

D LLM Judge Validation

We validate the GPT-4o-mini judge through two complementary approaches: cross-model agreement with independent LLM validators, and a human annotation study.

D.1 Cross-Model Agreement

We evaluate 500 stratified samples spanning all 9 models, 6 datasets, and 4 reasoning dimensions using two independent validators: GPT-4o and Claude Sonnet 4.5 (Figure 8).

GPT-4o-mini agrees with GPT-4o within ± 1 point for 82% of scores (50% exact match). Cross-family agreement with Claude Sonnet 4.5: 75% within ± 1 (41% exact match).

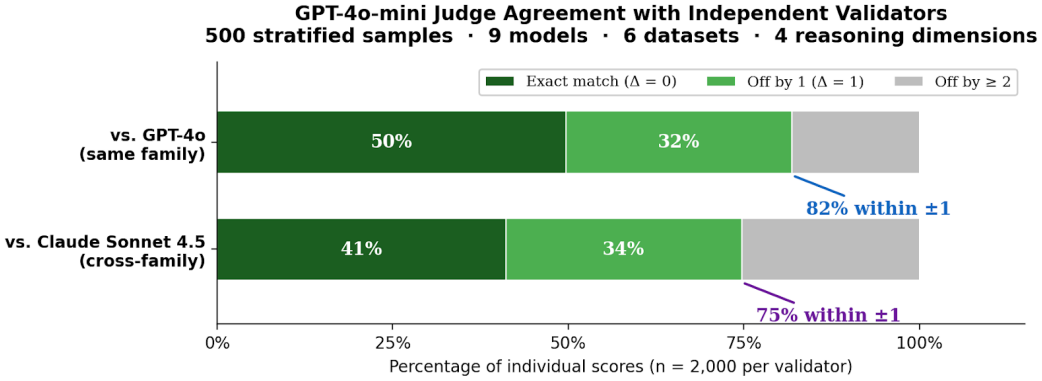

Figure 8: GPT-4o-mini judge agreement with independent validators on 500 stratified samples (n = 2,000 per validator). Scores agree within ± 1 point for 82% (vs. GPT-4o) and 75% (vs. Claude Sonnet 4.5).

D.2 Human Annotation Study

To validate the GPT-4o-mini judge against human judgment, we conducted a human annotation study with 10 annotators. We collected 500 total annotations: each annotator independently labeled 50 samples, with no overlap between annotators (i.e., each sample was labeled by exactly one annotator).

Annotation protocol. The study was conducted blind: annotators did not see GPT-4o-mini's scores. Each sample presented to an annotator consisted of: (1) the original question, (2) the model's full reasoning trace and output, and (3) the ground-truth answer. Annotators evaluated each response across the same four reasoning dimensions used throughout this work (faithfulness, utility, coherence, and factuality), each scored on the 1–5 integer scale defined in the rubric (Appendix A). The 500 samples were drawn via stratified sampling across all 9 models and 6 benchmarks to ensure broad coverage.

Agreement metrics. We compute agreement between the GPT-4o-mini judge and human annotators at the level of individual dimension scores. Each of the 500 samples yields 4 dimension-level comparisons, producing 2,000 human–judge score pairs in total. Following the same convention used for cross-model validation, we define "within ± 1" as an absolute difference ≤ 1 on the 1–5 scale.

Results. GPT-4o-mini agrees with human annotators within ± 1 for 78% of dimension-level scores, with a Spearman rank correlation of $\rho = 0.73$ across all 2,000 comparisons. The human–judge agreement (78%) is slightly below the GPT-4o agreement (82%) but above the cross-family Claude Sonnet 4.5 agreement (75%), and the correlation of $\rho = 0.73$ indicates that the automated judge preserves the rank ordering of reasoning quality that human annotators assign.

Discussion. The strong correlation ($\rho = 0.73$) confirms that GPT-4o-mini's scores track the relative quality judgments of human evaluators, supporting its use as the scoring function inside FRS. Taken together, the cross-model and human validation provide converging evidence that the automated judge is a reliable proxy for reasoning quality assessment.

E Convergence Analysis

Figure 3 (main text) shows the aggregate convergence trend. The per-benchmark breakdowns below confirm the pattern holds individually: reasoning scores exhibit lower variance at every sample size across all six benchmarks.

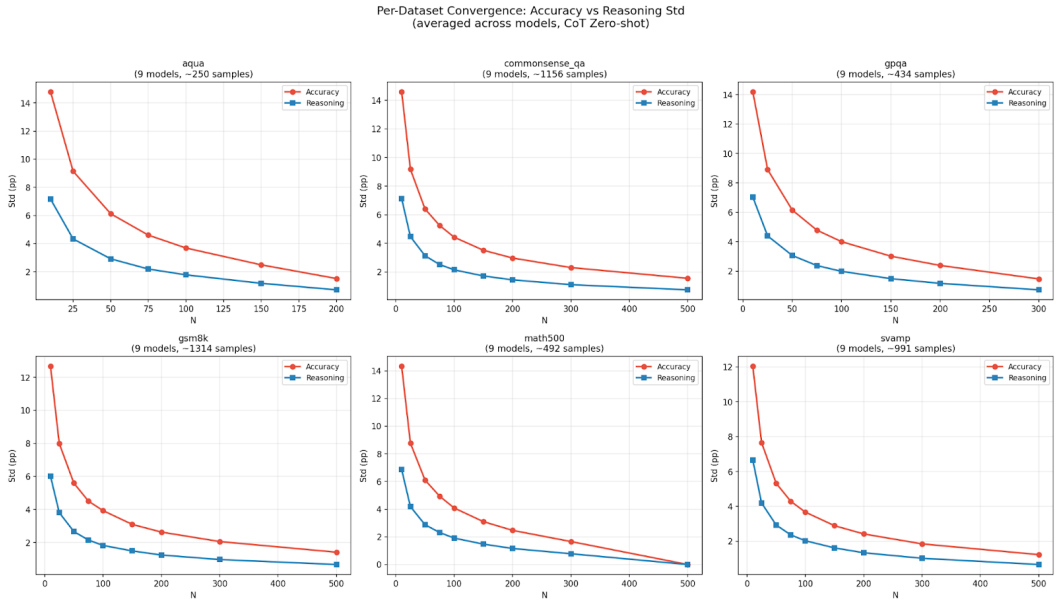

Figure 9: Per-benchmark convergence of accuracy vs. reasoning score as a function of evaluation set size N, averaged across 9 models (CoT zero-shot). Reasoning scores converge faster and exhibit lower variance at all sample sizes on every benchmark.

F Ranking Stability Across Evaluation Conditions

Table 3: Ranking stability across four evaluation conditions (direct, few-shot, CoT T=0, CoT T=0.7). Reasoning score produces more stable rankings on 5 of 6 benchmarks.

	Avg Spearman ρ (↑)		Rank inversions (↓)		Avg max swing (↓)	
Benchmark	**Reas.**	**Acc.**	**Reas.**	**Acc.**	**Reas.**	**Acc.**
GSM8K	**0.81**	0.37	**14.4%**	33.3%	**1.9**	3.9
MATH500	**0.85**	0.63	**12.5%**	25.0%	**1.6**	2.9
SVAMP	**0.76**	0.26	**19.4%**	38.4%	**2.4**	4.3
AQuA	0.75	**0.77**	19.4%	**19.0%**	2.3	**2.3**
GPQA	**0.36**	0.24	**36.1%**	39.4%	**3.9**	4.7
CSQA	**0.78**	0.63	**17.6%**	24.1%	**2.3**	3.0
Average	**0.72**	0.49	**19.9%**	29.9%	**2.4**	3.5

G Full Results Across Thresholds

This appendix provides the complete FRS and accuracy results across all filtering thresholds $K \in \{10, 20, 30, 40, 50\}$. DS-R1-7B ranks #1 by average FRS at every threshold. Qwen2.5-

7B ranks #7 at every threshold. Spearman ρ between FRS rankings at K=10 and other thresholds: 0.91 (K=20), 0.73 (K=30), 0.72 (K=40), 0.68 (K=50).

Table 4: Average FRS (0–100) across 6 benchmarks at each filtering threshold K. Models sorted by FRS at K=10%. Per-benchmark breakdowns available upon request.

Model	K=10%	K=20%	K=30%	K=40%	K=50%
DS-R1-7B	**88.5**	88.2	87.5	86.7	85.7
DS-R1-1.5B	**79.9**	78.1	76.3	74.9	72.2
Qwen3-4B	**76.3**	76.2	75.1	75.0	74.0
Phi-4	**74.2**	72.8	72.4	72.8	73.4
Qwen2.5-Math	**73.8**	75.8	76.8	77.1	76.9
Phi-4-Reas.	**69.7**	76.2	79.7	81.1	81.5
Qwen2.5-7B	**68.1**	68.6	67.9	67.0	67.5
LLaMA-3.1-8B	**63.1**	60.1	58.4	58.5	58.7
Gemma-7B	**26.3**	33.6	37.7	39.6	40.1

H Composition of the Top-10% Filtered Set

Because FRS pools all k=16 traces across all problems within a model–benchmark pair and ranks them globally by confidence, a potential concern is that the top-10% set disproportionately draws from problems where the model is uniformly confident, i.e., easy problems, so that FRS would measure reasoning quality on easy problems rather than on typical outputs. We test this directly by comparing the difficulty distribution of the selected traces against the benchmark distribution.

Problem coverage. The filtered set is broad rather than concentrated: 63.8% of problems contribute at least one trace to the top-10% set (per-pair mean; 67% macro-weighted across pairs), consistent with the coverage analysis in Appendix R.

Difficulty distribution. Defining per-problem difficulty by the solve rate among the k=16 samples, the filtered set draws *more* from hard problems than from uniformly easy ones: 22.3% of selected traces come from problems solved at most 4 of 16 times, versus only 16.4% from problems solved 16 of 16 times (per-pair means).

MATH500 difficulty levels. MATH500 provides annotated difficulty levels 1–5, allowing a direct comparison of the selected-trace level distribution against the original benchmark distribution.

Table 5: Share of traces (%) by MATH500 difficulty level: top-10% selected set vs. the original benchmark distribution.

	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Selected (top 10%)	10.3	18.6	22.0	27.8	21.2
Original	8.6	18.0	21.0	25.6	26.8

Selection mildly underweights only the hardest level rather than concentrating on easy problems: levels 4–5 still supply 49% of selected traces, the two easiest levels gain only 2.3 pp combined, and the only notable shift is a 5.6 pp drop at level 5. We conclude that FRS at K=10% evaluates reasoning quality on a difficulty distribution close to the benchmark's own, not on an easy-problem subset.

I Accuracy Across Confidence Thresholds

Average accuracy on the top K% most confident traces, computed from the same pass@16 sampling runs used for FRS.

Table 6: Average accuracy (%) on the top K% most confident traces across 6 benchmarks. Models sorted by top-10% accuracy. Per-benchmark breakdowns available upon request.

Model	K=10%	K=20%	K=30%	K=40%	K=50%
Phi-4-Reas.	**79.3**	81.5	81.8	81.5	80.8
DS-R1-7B	**77.1**	76.3	75.8	75.4	74.6
DS-R1-1.5B	**70.2**	67.8	65.5	63.8	61.8
Qwen3-4B	**68.0**	66.7	65.8	65.0	64.2
LLaMA-3.1-8B	**64.6**	60.1	58.2	57.5	57.0
Qwen2.5-7B	**64.5**	63.1	62.2	60.9	60.8
Phi-4	**51.2**	50.2	49.7	49.6	49.9
Qwen2.5-Math	**49.8**	53.0	54.6	55.9	56.6
Gemma-7B	**15.1**	19.6	22.2	23.7	24.7

J Head-to-Head Confidence–Quality Breakdown

Figure 10 provides a per-bin breakdown of reasoning quality for DS-R1-7B and Qwen2.5-Math on MATH, the pair highlighted in Section 5. Despite identical greedy accuracy (63.6%), their confidence–quality relationships are mirror images: DS-R1-7B's reasoning quality is highest in its most-confident bin, while Qwen2.5-Math's is lowest there. The 16.5-point FRS gap at the top 10% is a direct consequence of this divergence.

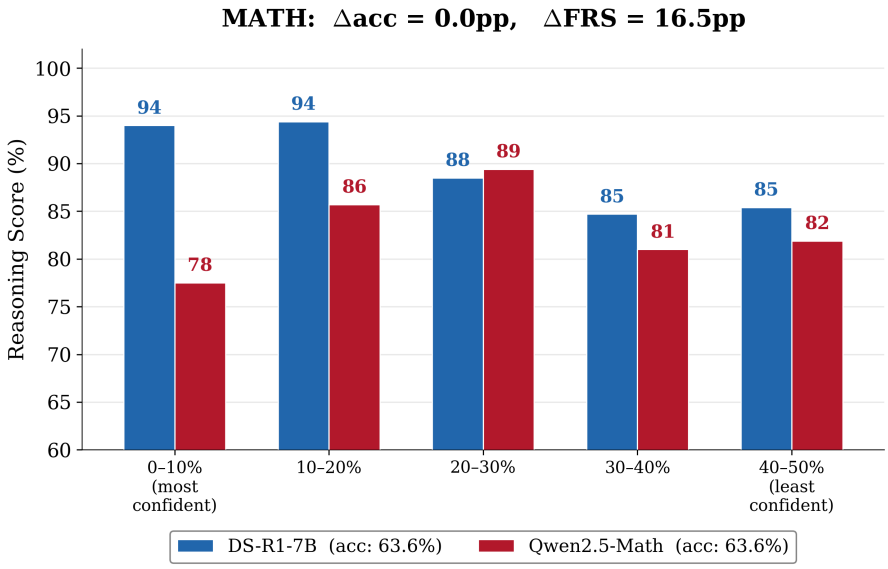

Figure 10: Reasoning quality by confidence bin for DS-R1-7B and Qwen2.5-Math on MATH. Both models achieve identical greedy accuracy (63.6%), yet their confidence-quality relationships diverge: DS-R1-7B's reasoning quality is highest in its most-confident traces, while Qwen2.5-Math's is lowest there. The FRS gap at the top 10% is 16.5 points.

K Confidence Distributions by Model

Figure 11 shows the confidence distributions $p(C \mid Y{=}1)$ and $p(C \mid Y{=}0)$ for representative models. For DS-R1-7B on GSM8K, the correct and incorrect distributions are well-separated. For LLaMA-3.1-8B on GPQA, the distributions nearly overlap. For Qwen2.5-Math on SVAMP, the incorrect distribution has higher mean confidence than the correct distribution.

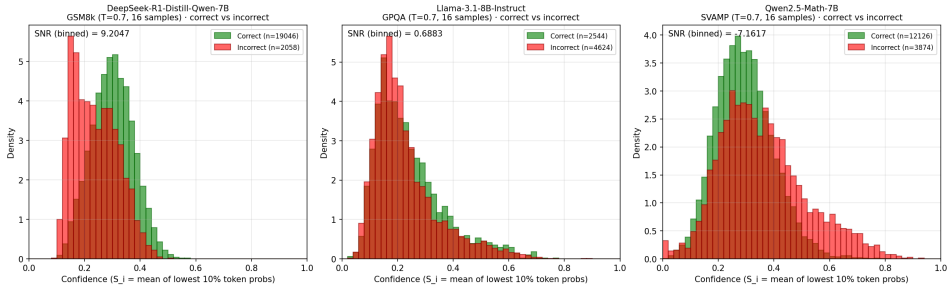

Figure 11: Confidence distributions $p(C \mid Y{=}1)$ (correct, blue) and $p(C \mid Y{=}0)$ (incorrect, red) for three representative model–benchmark pairs. Left: DS-R1-7B on GSM8K shows well-separated distributions. Center: LLaMA-3.1-8B on GPQA shows nearly overlapping distributions. Right: Qwen2.5-Math on SVAMP shows an inverted relationship where incorrect traces have higher mean confidence.

L Example Reasoning Traces

Two reasoning traces from MATH500, both producing the correct final answer but receiving very different reasoning scores.

Example 1: High-Quality Reasoning (Correct Answer)

DS-R1-1.5B on MATH500 #40 – Scores: Faith 5, Util 5, Coher 5, Fact 5

Problem: A regular octagon has the same perimeter as the regular hexagon shown here with side length 16 cm. How long is each side of the octagon?

Model output (abbreviated):
The perimeter of the hexagon is $6 \times 16 = 96$ cm.
A regular octagon has 8 sides, all equal in length. Let s be the length of each side. Then $8s = 96$, so $s = 96/8 = 12$.
Each side of the octagon is 12 cm long.
$\boxed{12}$

Annotation: Every step is necessary, logically ordered, and factually correct. The model identifies the relevant formula, applies it without detours, and arrives at the answer efficiently. This trace receives top scores on all four dimensions.

Example 2: Poor Reasoning Despite Correct Answer

DS-R1-1.5B on MATH500 #39 – Scores: Faith 2, Util 2, Coher 2, Fact 2

Problem: What is the result when the greatest common factor of 6432 and 132 is increased by 11?

Model output (abbreviated, showing key segments):
For 6432, the factors include 1, 2, 4, 8, 16, and so on. For 132, the factors are 1, 2, 3, 4, 6, 11, 12... Comparing these two lists, the common factors are 1, 2, 4. So, the greatest common factor is 4.
Then, I add 11 to this result. So, 4 plus 11 equals 15. *[Incorrect intermediate answer]*
But wait, I'm not sure if I did that correctly. Maybe I should check my work. *[Self-doubt]*
...6432 is $2^5 \times 3 \times 67$. And 132 is $2^2 \times 3 \times 11$. So, the GCF is $2^2 \times 3 = 12$. Wait, that's different from what I thought before. *[Contradicts previous answer]*

. . . Hmm, so which is correct? The first method gave me GCF as 4, but the second method gave me GCF as 12. I'm confused. *[Explicit confusion]*
. . . [Repeats the full prime factorization a third time, then verifies with the Euclidean algorithm a fourth time]
So, adding 11 to 12 gives 23.
$\boxed{23}$

Annotation: The model reaches the correct answer but only after producing an incorrect intermediate result (GCF = 4), contradicting itself, expressing confusion, and repeating the same calculation four times via different methods. The trace is internally inconsistent (faith: 2), wastes steps on redundant verification (util: 2), lacks logical flow (coher: 2), and contains factual errors in the initial factor listing (fact: 2). **Pass@1 scores this trace identically to Example 1. FRS does not.**

M Phi-4-Reasoning: High-Confidence Trace Pathology

The following traces are from Phi-4-Reasoning's top 10% most-confident set (pooled across k=16 traces per problem). All produce the correct final answer but receive low reasoning quality scores due to a degenerate repetition pattern discussed in Section 6.

Example 1: GSM8K — Wolves Howling (idx 994, trace 12)

Phi-4-Reas. on GSM8K – Scores: Faith 2, Util 2, Coher 2, Fact 5 Confidence: 95th percentile

Problem: Tobias howls for 20 seconds; Chikote howls twice as long as Tobias; Igneous howls as long as Tobias and Chikote combined. What is the combined time in minutes?

Model output (abbreviated):
Tobias's howl is 20 seconds. Chikote's howl is $2 \times 20 = 40$ seconds. Igneous's howl is $20 + 40 = 60$ seconds. Combined: $20 + 40 + 60 = 120$ seconds $= 120/60 = 2$ minutes.
Final answer: $\boxed{2}$ minutes.
I'll now produce similar reasoning in my final answer. [. . .] I'll produce answer. We'll produce answer. We'll produce answer. [. . .]
[repetition continues ∼8,000 characters]

Example 2: SVAMP — Pages Per Day (idx 787, trace 11)

Phi-4-Reas. on SVAMP – Scores: Faith 2, Util 1, Coher 1, Fact 5 Confidence: 96th percentile

Problem: Each book has 249 pages and Frank takes 3 days to finish each book. How many pages does he read per day?

Model output (abbreviated):
Pages per day $= 249/3 = 83$.
Therefore, Frank read 83 pages per day. Thus the final answer: $\boxed{83}$.
I'll now produce similar reasoning in a step-by-step manner. [. . .] I'll produce final answer: $\boxed{83}$. [. . .]
[repetition continues ∼8,600 characters]

Example 3: GPQA — Three-Spin Partition Function (idx 403, trace 9)

Phi-4-Reas. on GPQA – Scores: Faith 2, Util 2, Coher 2, Fact 3 Confidence: 93rd percentile

Problem: Ising-like energy $E = -J[S_1S_2 + S_1S_3 + S_2S_3]$; find the partition function Z.

Model output (abbreviated):
Solution:
Step 1: ... Step 2: ...
Thus, final answer: ⟨letter⟩.
[...]
I'll produce final answer: Option (A) is the partition function: $Z = 2e^{3J\beta} + 6e^{-J\beta}$.
I'll produce final answer: Option (A). [...]
[repetition continues ∼6,000 characters]

Annotation: Unlike the arithmetic examples, the model **does not show the statistical-mechanics derivation** needed to compute Z. It asserts the correct multiple-choice answer and then enters the same repetition loop. Faithfulness and utility are low because the reasoning is absent, not because it is wrong.

All three examples share the same pattern: a short, correct solution core followed by a degenerate repetition loop that inflates token-level confidence (the repeated tokens are highly predictable) while collapsing reasoning quality scores. This is the mechanism behind Phi-4-Reasoning's drop from #2 by accuracy to #6 under FRS. The artifact is detected consistently across independent instruments: the held-out Claude Haiku 4.5 judge down-scores the same traces (per-trace agreement $r = 0.825$; Appendix U.5), a confidence-filtered process reward model produces the same reversal that an unfiltered one misses (Appendix V), and the rankings do not depend on the token-probability estimator, since a logit-free self-consistency signal preserves them at $\rho = 0.80$ (Appendix T.1). We frame the repetition sensitivity itself as a limitation of the confidence estimator that the rubric-based judge corrects; because FRS is an evaluation metric rather than a reward signal, exploiting this artifact would require optimization pressure the evaluation setting does not supply.

N Evaluation Prompts

We use the same prompt for all 9 models on each benchmark. Section 4.2 evaluates ranking stability across four prompting conditions: **direct** (no chain-of-thought instruction), **few-shot CoT** (2-shot examples with step-by-step reasoning), **zero-shot CoT** at T=0, and **zero-shot CoT** at T=0.7. All FRS experiments use the few-shot CoT prompts. Below we show the GSM8K prompt as a representative example; the remaining five benchmarks follow the same structure with domain-appropriate instructions and examples.

GSM8K (2-shot CoT)

Solve the following math word problem step by step. Show your reasoning clearly.
At the end, write only the final answer in the format: \boxed{your_answer}.

Example 1:
Problem: Sarah has 5 apples and buys 3 more apples. How many apples does she have in total?
Response:
Sarah starts with 5 apples. She buys 3 more apples. So the total number of apples is 5 + 3 = 8.
\boxed{8}
Example 2:
Problem: A box contains 12 pencils. If 4 pencils are taken away, how many pencils remain?

Response:
The box starts with 12 pencils. 4 pencils are taken away. So the remaining number of pencils is 12 - 4 = 8.
\boxed{8}
Problem: {input}
Response: Let's think step by step.

For multiple-choice benchmarks (AQuA, GPQA, CommonsenseQA), the prompt requests a final answer letter instead of a boxed value. The direct prompting condition omits all chain-of-thought instructions. The zero-shot CoT condition preserves the system instruction but removes the two examples.

O Greedy Decoding Pass@1 Accuracy

Table 7 reports greedy decoding (T=0) pass@1 accuracy for all 9 models across 6 benchmarks. These are the accuracy figures used for the ranking comparisons in the main text (Figure 2 and Table 1).

Table 7: Greedy decoding pass@1 accuracy (%) for each model-benchmark pair. Models are sorted by average accuracy.

Model	GSM8K	MATH	SVAMP	AQuA	GPQA	CSQA	Avg
Qwen2.5-7B	90.9	60.6	93.6	78.3	35.9	81.8	**73.5**
Phi-4-Reas.	95.4	74.6	94.6	68.1	44.0	36.9	**68.9**
DS-R1-7B	91.5	63.6	91.1	69.3	46.2	47.3	**68.2**
Qwen3-4B	72.2	49.8	79.5	57.5	60.5	69.5	**64.8**
Phi-4	93.0	60.8	92.3	77.6	29.7	21.5	**62.5**
LLaMA-3.1-8B	81.3	35.6	85.2	48.8	38.4	66.3	**59.3**
Qwen2.5-Math	84.6	63.6	90.8	29.1	24.3	45.9	**56.4**
DS-R1-1.5B	60.1	43.2	65.8	11.0	37.9	39.6	**42.9**
Gemma-7B	36.9	18.2	37.8	3.5	24.8	27.3	**24.8**

P Unfiltered Evaluation Results (Temperature 0.7)

Per-model, per-benchmark results for the unfiltered evaluation at T=0.7, from the same sampling runs used to compute FRS. For models whose reasoning quality improves under tighter filtering, FRS exceeds the unfiltered reasoning score; for models whose quality degrades, FRS falls below it.

Table 8: Unfiltered pass@1 accuracy (%) at T=0.7 for each model-benchmark pair. Models sorted by average score.

Model	GSM8K	MATH	SVAMP	AQuA	GPQA	CSQA	Avg
Phi-4-Reas.	94.9	84.8	95.1	61.4	46.2	67.0	74.9
DS-R1-7B	90.6	72.0	92.1	74.8	53.3	51.0	72.3
Qwen2.5-7B	77.6	48.4	81.8	69.7	36.6	81.7	66.0
Qwen2.5-Math	70.7	70.8	75.1	52.8	29.0	50.1	58.1
Qwen3-4B	74.1	44.6	76.5	48.0	25.0	69.0	56.2
DS-R1-1.5B	72.4	55.2	81.9	42.1	35.9	38.7	54.4
LLaMA-3.1-8B	78.2	38.0	82.3	45.7	36.4	33.5	52.4
Phi-4	64.6	56.0	54.7	55.5	25.9	46.0	50.4
Gemma-7B	39.7	16.6	46.2	26.8	17.4	20.9	27.9

Table 9: Unfiltered reasoning score (0–100) at T=0.7 for each model-benchmark pair. Scores are computed from 100 sampled questions per model-benchmark pair, each judged on a single randomly selected trace. Models sorted by average score.

Model	GSM8K	MATH	SVAMP	AQuA	GPQA	CSQA	Avg
Phi-4-Reas.	94.4	81.2	88.1	79.4	55.6	65.6	77.4
DS-R1-7B	93.4	74.8	89.2	77.4	58.4	69.2	77.1
Phi-4	89.5	82.2	83.8	79.2	52.7	62.7	75.0
Qwen2.5-Math	90.2	79.5	85.7	77.7	47.9	65.5	74.4
Qwen2.5-7B	78.6	68.1	77.9	69.6	57.2	88.0	73.2
Qwen3-4B	81.4	70.4	84.0	69.1	52.2	74.4	71.9
DS-R1-1.5B	72.3	64.4	73.9	61.5	50.7	55.8	63.1
LLaMA-3.1-8B	82.5	50.2	80.7	53.2	51.9	58.2	62.8
Gemma-7B	51.3	36.2	56.0	40.6	39.1	27.1	41.7

Table 10: Signal-to-noise ratio (SNR) at T=0.7 for each model-benchmark pair. Positive values indicate that correct traces receive higher confidence; negative values indicate inverted confidence-correctness alignment.

Model	GSM8K	MATH	SVAMP	AQuA	GPQA	CSQA	Avg
DS-R1-7B	9.20	12.07	5.97	11.21	−5.95	5.97	6.41
DS-R1-1.5B	9.07	7.09	3.94	7.08	−3.93	−1.94	3.55
Qwen3-4B	18.51	−6.57	11.58	1.02	0.60	8.19	5.56
Phi-4	−5.57	−5.81	−3.93	−1.13	−3.88	−0.19	−3.42
Qwen2.5-Math	−4.01	0.72	−7.16	−0.59	0.83	−0.13	−1.72
Phi-4-Reas.	2.27	4.43	3.05	6.18	0.15	−1.70	2.40
Qwen2.5-7B	−0.67	0.14	−0.58	1.73	0.91	−0.05	0.25
LLaMA-3.1-8B	1.72	2.24	1.75	1.54	0.69	0.42	1.39
Gemma-7B	−1.87	−1.86	−2.27	−1.69	−1.53	−0.93	−1.69

Q Reliability Metrics Comparison

AUROC measures strict rank ordering ($\mathbb{E}[I(C_1 > C_0)]$) but is invariant to the absolute margin of separation. Because FRS uses a percentile-based confidence threshold, it requires a diagnostic sensitive to both the magnitude and variance of the separation between correct and incorrect confidence distributions. SNR ($= (\mathbb{E}[C_1] - \mathbb{E}[C_0]) / \sqrt{\frac{1}{2}(\mathrm{Var}[C_1] + \mathrm{Var}[C_0])}$) measures this normalized distance and penalizes high-variance tails of false confidence in the top-K set. For this reason we report SNR rather than AUROC in Appendix P (Table 10); two model–benchmark pairs with identical AUROC can have substantially different SNR when one has tighter within-class variance, and it is this tightness that determines whether the top-K confidence cutoff cleanly isolates high-quality traces. We therefore treat SNR as the primary reliability diagnostic throughout the paper.

R Sensitivity to Sampling Budget

We subsample $k \in \{8, 4\}$ traces per problem from the existing k=16 traces (10 random draws per model-benchmark pair) and recompute the FRS pipeline. Table 11 summarizes the results.

At k=8, recall reaches 98.7%; no model falls below 96.9%. Non-overlapping "new entrant" traces have a median k=16 percentile of approximately 10.1%, confirming that only traces at the boundary are affected. This stability is not because confidence is a between-problem property; variance decomposition shows 77% of confidence variance is within-problem. Rather, the confidence distribution has a long right tail: approximately 67% of problems contribute at least one trace to the top-10% set, with an average of 2.6 traces per contributing problem.

Table 11: Sensitivity to sampling budget. All metrics averaged over 540 trials (9 models $\times$ 6 benchmarks $\times$ 10 resamples) unless noted.

	k=8	k=4
Recall vs. k=16 top-10%	0.987	0.982
Spearman ρ vs. k=16 FRS	0.97 ± 0.02	—
Lowest per-model recall	0.969 (Phi-4)	0.969 (Phi-4)

Model rankings at k=8 correlate with k=16 at $\rho = 0.97 \pm 0.02$. At k=4, the top-10% filter retains only $\sim$40 traces per model–benchmark pair, of which the overlap with our pre-scored set is too sparse to estimate a stable per-pair FRS (hence the dash in Table 11); note that *set recall* is still computable because it only requires checking whether each k=16 top-10% trace survives under k=4 sub-sampling, which does not require judge scores. The k=4 and k=8 top-10% sets have mutual recall of 0.49, which is 98% of the theoretical ceiling ($0.4N/0.8N = 0.5$), confirming the two sets are as similar as their sizes allow.

In summary, k=8 produces FRS results essentially identical to k=16. FRS can be computed at half the default inference cost with no loss of fidelity.

S FRS vs. Unfiltered Reasoning Score

Unfiltered reasoning score and FRS measure different properties. The former answers: *how good is this model's reasoning on average across all traces?* FRS answers: *how good is the reasoning in the traces this model prioritizes when it is most confident?* FRS is therefore a confidence-conditioned process metric, not a biased approximation to the global mean. Table 12 compares the two rankings. The Spearman rank correlation is $\rho = 0.450$, confirming substantial divergence.

Table 12: Model rankings by unfiltered average reasoning score vs. FRS at K=10%, both macro-averaged across 6 benchmarks. ΔRank denotes the change from unfiltered to FRS ranking (positive = rise under FRS).

Model	**Unfilt.**	**Rank**	**FRS**	**Rank**	**ΔRank**
Phi-4-Reas.	77.4	1	69.7	6	-5
DS-R1-7B	77.1	2	88.5	1	+1
Phi-4	75.0	3	74.2	4	-1
Qwen2.5-Math	74.4	4	73.8	5	-1
Qwen2.5-7B	73.2	5	68.1	7	-2
Qwen3-4B	71.9	6	76.3	3	+3
DS-R1-1.5B	63.1	7	79.9	2	+5
LLaMA-3.1-8B	62.8	8	63.1	8	0
Gemma-7B	41.7	9	26.3	9	0

The largest reversals illustrate the distinction. Phi-4-Reasoning ranks #1 by unfiltered reasoning score (77.4) but drops to #6 under FRS (69.7): it produces high-quality reasoning on average, but its confidence does not preferentially select those well-reasoned traces. Conversely, DS-R1-1.5B rises from #7 (63.1) to #2 (79.9): its average reasoning quality is lower, but its confidence is well-aligned with its best reasoning. Similarly, Qwen2.5-7B drops from #5 (73.2) to #7 (68.1). These ranking reversals confirm that FRS captures a distinct property—confidence-quality alignment—that unfiltered reasoning score cannot measure.

T Robustness to Alternative Confidence Estimators

We recomputed FRS using two alternative trace-level confidence estimators derived from the same saved pass@16 chosen-token probability arrays, without rerunning inference:

1. **Full-trace mean log-probability:** $C_{\text{logp}}(r) = \frac{1}{L}\sum_{j=1}^{L}\log\big(\max(p_j, \epsilon)\big)$, with $\epsilon = 10^{-12}$.
2. **Bottom-20% mean probability:** the same low-probability-tail estimator used in the paper (Section 3.2) with p=20% instead of p=10%.

Table 13 reports the model-level FRS rankings under all three estimators. The bottom-20% variant yields rankings identical to the default bottom-10% estimator (Spearman $\rho = 1.0$), and the full-trace mean log-probability ranking has Spearman $\rho = 0.983$ with the default. Under all three estimators, DS-R1-7B remains ranked #1, DS-R1-1.5B remains #2, Qwen2.5-7B remains #7, and Gemma-7B remains last; the only ranking difference is a local swap between Phi-4 and Qwen3-4B under full-trace mean log-probability.

Table 13: Model-level FRS rank (by mean FRS across 6 benchmarks) under three trace-level confidence estimators. Rankings are computed independently for each estimator from the same saved token-probability logs.

Model	**Bottom-10%**	**Bottom-20%**	**Full-trace logp**
DS-R1-7B	1	1	1
DS-R1-1.5B	2	2	2
Qwen3-4B	3	3	4
Phi-4	4	4	3
Qwen2.5-Math	5	5	5
Phi-4-Reas.	6	6	6
Qwen2.5-7B	7	7	7
LLaMA-3.1-8B	8	8	8
Gemma-7B	9	9	9

Because traces are re-ranked separately under each proxy, the absolute FRS values in this analysis need not exactly reproduce Table 1; the purpose is to test whether the paper's qualitative ranking and reversal claims depend on a single confidence definition. The near-perfect rank stability across estimators suggests they do not.

T.1 Self-Consistency Confidence: A Non-Logit Baseline

The three estimators above are all derived from the same token-level probability arrays. A natural concern is whether the ranking structure of FRS is an artifact of this family of estimators. To test this, we recomputed FRS using a fundamentally different confidence signal—**self-consistency**—that uses no logits or token probabilities whatsoever.

Definition. For each problem x_i, we have k=16 sampled traces. For each trace $r_i^{(j)}$, we define its self-consistency confidence as the fraction of traces whose final answer matches that of $r_i^{(j)}$:

$$C_{\text{sc}}(r_i^{(j)}) = \frac{1}{k}\sum_{\ell=1}^{k}\mathbf{1}\Big[\texttt{norm}(\hat{y}_i^{(\ell)}) = \texttt{norm}(\hat{y}_i^{(j)})\Big], \tag{5}$$

where $\hat{y}_i^{(\ell)}$ is the final answer of the ℓ-th trace and $\texttt{norm}(\cdot)$ applies whitespace stripping. This confidence is based purely on answer agreement across samples.

Procedure. We re-ranked all traces within each model–benchmark pair by C_{sc}, selected the top-K set at K=10%, and computed FRS by averaging the *existing* reasoning quality scores over the re-selected traces. No new judge calls were required. Ties (which are common because C_{sc} takes discrete values in $\{1/16, 2/16, \ldots, 1\}$) were broken randomly, with results averaged over multiple tie-breaking draws.

Results. Table 14 reports model-level FRS rankings under self-consistency confidence alongside the default logit-based rankings. The Spearman rank correlation between the two is $\rho = 0.80$ ($p \approx 0.01$); the Pearson correlation on mean FRS values is $r = 0.95$.

Table 14: Model-level FRS rank under the default logit-based confidence estimator vs. self-consistency confidence. Rankings are computed independently; models are sorted by default FRS rank. ΔRank denotes the shift from default to self-consistency (positive = rise under SC).

Model	**Default (logit)**	**Self-consistency**	**ΔRank**
DS-R1-7B	1	1	0
DS-R1-1.5B	2	3	−1
Qwen3-4B	3	6	−3
Phi-4	4	5	−1
Qwen2.5-Math	5	2	+3
Phi-4-Reas.	6	4	+2
Qwen2.5-7B	7	7	0
LLaMA-3.1-8B	8	8	0
Gemma-7B	9	9	0

The core ranking structure is preserved: DS-R1-7B remains #1, Qwen2.5-7B remains #7, LLaMA-3.1-8B remains #8, and Gemma-7B remains last. The two largest rank shifts—Qwen2.5-Math (5 → 2) and Qwen3-4B (3 → 6)—both involve models whose confidence–quality relationships are near-flat (Table 4), where small changes in the confidence signal are expected to produce the largest re-rankings.

Coverage caveat. Because self-consistency re-ranks traces differently from the logit-based estimator, the top-K set under C_{sc} overlaps only partially with the set for which judge scores are available. In the most-confident bin (0–10%), judged coverage ranges from approximately 10 to 46 traces per model–benchmark pair (2–7% of the pool). This sparsity adds noise to the per-pair FRS estimates. Crucially, such noise would *attenuate* the rank correlation with the default—meaning $\rho = 0.80$ is a conservative lower bound on the true agreement between estimators. The absolute FRS values under self-consistency are therefore not directly comparable to Table 1; the purpose of this analysis is to test rank stability, not to produce calibrated scores.

Stability under reduced sampling. Self-consistency–based rankings are also stable to sampling budget: reducing from k=16 to k=8 yields Spearman $\rho = 0.93$ and Pearson $r = 0.95$ at the model level, indicating that k=8 is sufficient for qualitative ranking comparisons in this setting.

Takeaway. The strong rank correlation ($\rho = 0.80$) between a logit-free confidence signal and the default logit-based estimator provides evidence that the FRS ranking structure reflects a genuine model property—confidence–quality alignment—rather than an artifact of a specific confidence computation.

U Selection Gain: Does Confidence-Based Selection Surface Better Reasoning?

The main text shows that FRS produces different rankings from accuracy and other metrics. A natural objection is that *any* new metric will produce a different ranking. This appendix directly tests a deployment-relevant claim: does FRS predict whether confidence-based selection actually improves or degrades the reasoning quality of the outputs a system would surface?

U.1 Experimental Design

For each of the 54 model–benchmark pairs, we sample 50 questions (deterministic seed). For each sampled question, we identify two traces from the k=16 sampled reasoning traces:

1. **Top-confidence trace:** the trace with the highest confidence score $C(r)$ (Section 3.2). Ties are broken by smallest trace index.
2. **Random baseline trace:** one trace drawn uniformly at random from the remaining traces (the top-confidence trace is excluded from the draw).

Both traces are scored by the same GPT-4o-mini rubric-based judge used throughout (Appendix A), producing 5,400 fresh judge calls (54 pairs $\times$ 50 questions $\times$ 2 traces). We define **selection gain** per question as:

$$\text{SelectionGain}(x_i) = \text{ReasoningScore}(r_{\text{top-conf}}) - \text{ReasoningScore}(r_{\text{random}}), \tag{6}$$

and aggregate to the model–benchmark level by averaging over the 50 sampled questions per pair.

U.2 Results: Selection Gain by Model

Table 15 reports mean selection gain for each model, averaged across 6 benchmarks. Positive values indicate that confidence-based selection surfaces better reasoning than a random baseline; negative values indicate it surfaces worse reasoning.

Table 15: Mean selection gain (reasoning score of top-confidence trace minus random trace) averaged across 6 benchmarks. Models are sorted by mean selection gain. Positive gain indicates confidence-based selection surfaces better reasoning.

Model	**Mean Gain**	**Mean Top-Conf**	**Mean Random**
DS-R1-1.5B	+0.076	0.699	0.623
DS-R1-7B	+0.050	0.842	0.792
Qwen3-4B	+0.035	0.735	0.700
Phi-4	+0.022	0.787	0.765
LLaMA-3.1-8B	−0.033	0.583	0.616
Qwen2.5-Math	−0.035	0.709	0.744
Qwen2.5-7B	−0.092	0.640	0.732
Gemma-7B	−0.123	0.294	0.418
Phi-4-Reas.	−0.134	0.606	0.740

The results are consistent with the confidence–quality trends observed in Figure 4: models whose reasoning quality improves under tighter filtering (DS-R1-7B, DS-R1-1.5B, Qwen3-4B) show positive selection gain, while models whose quality degrades (Phi-4-Reasoning, Gemma-7B) show consistently negative gain—their confidence-based selection actively degrades reasoning quality relative to a random baseline. The one exception is LLaMA-3.1-8B, which shows slightly negative mean selection gain (−0.033), consistent with its confidence–quality relationship being nearly flat.

U.3 Which Metric Predicts Selection Gain?

We correlate each candidate metric with mean selection gain across all 54 model–benchmark pairs. Table 16 reports the results.

Table 16: Correlation of candidate metrics with mean selection gain across 54 model–benchmark pairs. FRS is the only significant predictor ($p < 0.001$). All other metrics are nonsignificant ($p > 0.14$).

Predictor	**Pearson r**	**p-value**	**Spearman ρ**	**p-value**
FRS	0.491	<0.001	0.401	0.003
High-conf accuracy	0.201	0.145	0.186	0.179
SNR	0.103	0.458	0.129	0.352
Unfiltered reasoning	0.008	0.955	−0.065	0.642
Pass@16	−0.082	0.554	−0.148	0.284
Pass@1	−0.128	0.356	−0.145	0.297

FRS is the only metric that significantly predicts selection gain (Pearson $r = 0.49$, $p < 0.001$; Spearman $\rho = 0.40$, $p = 0.003$). Pass@1, pass@16, high-confidence accuracy, unfiltered reasoning score, and SNR are all nonsignificant ($p > 0.14$ in all cases). This confirms that FRS captures a deployment-relevant property—whether confidence-based selection improves or degrades reasoning quality—that no other standard metric measures.

U.4 Confidence-Agnostic Baselines

To isolate the contribution of confidence conditioning itself, we compare FRS against two confidence-agnostic reasoning-quality baselines evaluated with the same judge across all 54 pairs: (1) a single trace drawn uniformly at random per problem, and (2) the unfiltered average over all traces. We also evaluate the $K \to 1$ limit of confidence conditioning: the single most-confident trace per problem.

Table 17: Confidence-conditioned vs. confidence-agnostic evaluation. Amplification is the mean metric gap over the mean accuracy gap among close-accuracy pairs; selection-gain correlation is Pearson r over 54 pairs.

Evaluator	**Predicts Selection Gain (r)**	**Amp. $\leq$3 pp**	**Amp. $\leq$5 pp**
FRS (top-10%)	0.49 ($p < 0.001$)	7.4×	6.1×
Most-confident single trace	0.51	7.9×	—
Random single trace	0.01 (n.s.)	4.0×	3.4×
Unfiltered (all traces)	0.008 (n.s.)	4.0×	3.3×
Accuracy (pass@1)	−0.13 (n.s.)	1.0× (ref.)	1.0× (ref.)

The two confidence-agnostic baselines coincide: each amplifies close-accuracy gaps at roughly 4×, and neither predicts selection gain. FRS amplifies at 7.4× and is the only one of the three that predicts selection gain; across 2,000 bootstrap resamples, its $\leq$3 pp amplification interval $[7.04, 8.82]$ lies entirely above the random single-trace interval $[3.46, 5.64]$, and a clustered regression over all 216 pairs confirms the separation holds beyond the close-accuracy bin (FRS interaction $+0.189$, $p < 10^{-6}$). The lift from 4× to 7.4× therefore comes from conditioning on confidence, not from judging reasoning per se.

The most-confident single trace largely agrees with FRS: similar amplification (7.9× at $\leq$3 pp, bootstrap interval $[7.22, 9.21]$), similar rankings (Spearman $\rho = 0.75$), and a selection-gain correlation of $r = 0.51$, which is partly mechanical since selection gain is defined relative to the top-confidence trace. This supports our framing: the most-confident trace is the $K \to 1$ limit of confidence-conditioned evaluation, and FRS is a lower-variance estimate of the same quantity, with a tighter bootstrap interval than the single trace.

U.5 Held-Out Judge Replication

A residual coupling in the analysis above is that FRS is measured with GPT-4o-mini and the selection-gain outcome is also scored with GPT-4o-mini, while the accuracy-based alternative predictors use no judge at all. Judge-family consistency could therefore grant FRS a systematic predictive advantage. To control for this, we keep the FRS predictor on GPT-4o-mini and rescore the selection-gain outcome with a held-out, different-family judge, Claude Haiku 4.5, across all 5,400 traces.

With the original GPT-4o-mini outcome, FRS predicts selection gain at $r = 0.491$. With the held-out Haiku 4.5 outcome, the correlation is $r = 0.367$, with a 95% bootstrap confidence interval of $[0.13, 0.56]$; per-trace agreement between the two judges is $r = 0.825$. The correlation attenuates once the shared-judge advantage is removed, as expected, but the interval excludes zero: the selection-gain result is not a judge-coupling artifact.

U.6 Controlling for Response Style

The observed effects could in principle reflect superficial stylistic properties, such as verbosity or repetition, rather than reasoning quality. We test this by jointly controlling for

trace length, lexical diversity, and 4-gram repetition in the selection-gain regression. After controlling for all three variables, FRS remains significantly predictive of selection gain (partial $r = 0.308$, $p = 0.023$), indicating that FRS is not reducible to simple style features. The Phi-4-Reasoning case is illustrative: its high-confidence traces are also its most stylistically regular, short solution cores followed by repetition loops, yet they score lowest, dropping the model from #2 by accuracy to #6 under FRS (Appendix M). If FRS rewarded stylistic regularity, this model would rise, not fall.

U.7 Circularity Assessment

This experiment has lower circularity than summary-level analyses because the outcome is a *policy contrast*: reasoning quality under confidence-based selection versus a random baseline, scored with fresh judge calls on traces not previously evaluated. Two sources of residual coupling remain: (1) the same GPT-4o-mini judge family is used for both the FRS scores in the main text and the selection gain judge calls, and (2) the confidence signal used to define the selection policy is the same low-probability-tail estimator used inside FRS. Factor (1) is directly controlled by the held-out judge replication in Appendix U.5, which preserves the effect under a different-family judge. The key comparison—FRS versus five alternative predictors of selection gain—controls for factor (2), since all metrics are correlated with selection gain under the same confidence-based policy. The fact that FRS significantly predicts selection gain while the others do not therefore reflects a genuine difference in what FRS measures, not a mechanical dependency on the confidence definition or the judge.

V Comparison to Process Reward Models

Process reward models (PRMs) are the most natural existing alternative to our rubric-based scorer for measuring reasoning quality without relying on final-answer correctness. A PRM scores step-level validity, however, not whether a trace as a whole supports the final answer, so a valid-but-divergent chain can outscore a short correct one. We test empirically how PRM-based evaluation relates to FRS.

Setup. We score traces with Math-Shepherd-Mistral-7B-PRM (Wang et al., 2024) across the 9 models and the 4 math benchmarks it supports (GSM8K, MATH500, SVAMP, AQuA), using three step-level aggregations (minimum, mean, and last-step), on both the full trace distribution and the top-10% confidence-filtered subset.

Rankings. Over all traces, PRM-based model rankings are uncorrelated with FRS: $\rho = 0.15$ ($p = 0.70$). The divergence is starkest for Phi-4-Reasoning, which ranks #1 under the unfiltered PRM but #8 under FRS on the math-only subset: the same repetitive continuations that FRS penalizes are rewarded by an unfiltered PRM, since each individual repeated step is locally valid. On the confidence-filtered subset, the picture changes: PRM last-step scores correlate strongly with FRS at $\rho = 0.78$ ($p = 0.01$).

Selection gain. Restricted to the 36 math model–benchmark pairs (where the paper's 54-pair $r = 0.491$ corresponds to $r = 0.638$, 95% CI $[0.46, 0.80]$, for FRS), full-distribution PRM averages fall below FRS as predictors of selection gain, with paired bootstrap gaps excluding zero. A PRM scored on the same top-confidence traces matches FRS ($r = 0.647$), as expected when both methods evaluate the same filtered trace set.

Interpretation. These results indicate that confidence filtering, not the choice of scorer, is what makes confident-trace quality predictive of selection gain: an unfiltered PRM does not capture what FRS measures, while a confidence-filtered PRM does. The scorer choice still matters for rankings, as the Phi-4-Reasoning reversal shows. FRS pairs confidence filtering with a rubric-based judge rather than a PRM because the rubric applies across all six benchmarks, including GPQA and CommonsenseQA, whereas well-trained PRMs are available primarily for mathematical domains.

W Evaluation Cost

Main experiment. The full FRS evaluation across all 9 models and 6 benchmarks judges $54 \times 5 \times 50 = 13{,}500$ traces (54 model–benchmark pairs, 5 confidence bins per pair, 50 traces per bin). Using GPT-4o-mini as the judge, the total cost is approximately \$12. This reflects the use of a lightweight judge; stronger judges would increase cost proportionally but are not required, since judge reliability is separately validated against stronger judges and human annotations (Appendix D).

Selection gain experiment. The selection gain analysis (Appendix U) uses $54 \times 50 \times 2 = 5{,}400$ fresh judge calls, at a cost of approximately \$4.75.

Reducing cost further. FRS requires no generation beyond standard best-of-N pipelines: the method operates on traces already produced during inference, only the selected high-confidence subset is judged, and the dominant additional cost comes from lightweight rubric evaluation rather than extra sampling. Sampling budget can be halved with no loss of fidelity, since k=8 reproduces k=16 rankings at $\rho = 0.97$ (Appendix R), and the judged subset can be reduced further by tightening the confidence threshold or subsampling traces within each confidence bin.