Reliable Chain-of-Thought via Prefix Consistency

Naoto Iwase[1] **Yuki Ichihara**[2,3] **Mohammad Atif Quamar**[3] **Junpei Komiyama**[3,4]
[1]Nagoya University [2]Nara Institute of Science and Technology
[3]Mohamed bin Zayed University of Artificial Intelligence [4]RIKEN AIP
naoto@iwase.dev {yuki.ichihara, mohammad.atif}@mbzuai.ac.ae junpei@komiyama.info
Project Page: https://naoto-iwase.github.io/prefix-consistency-page

Abstract

Large language models often improve accuracy on reasoning tasks by sampling multiple chain-of-thought (CoT) traces and aggregating them with majority voting, a test-time technique called self-consistency. When we truncate a CoT partway through and regenerate the remainder, we observe that traces with correct answers reproduce their original answer more often than traces with wrong answers. We use this difference as a reliability signal, **prefix consistency**, that weights each candidate answer by how often it reappears under regeneration. It requires no access to token log-probabilities or self-rating prompts. Across five reasoning models and four math and science benchmarks, prefix consistency is the best correctness predictor in most settings, and reweighting votes by this signal reaches the plateau accuracy of majority voting with up to $21\times$ fewer tokens (median $4.6\times$). Our code is available at https://github.com/naoto-iwase/prefix-consistency.

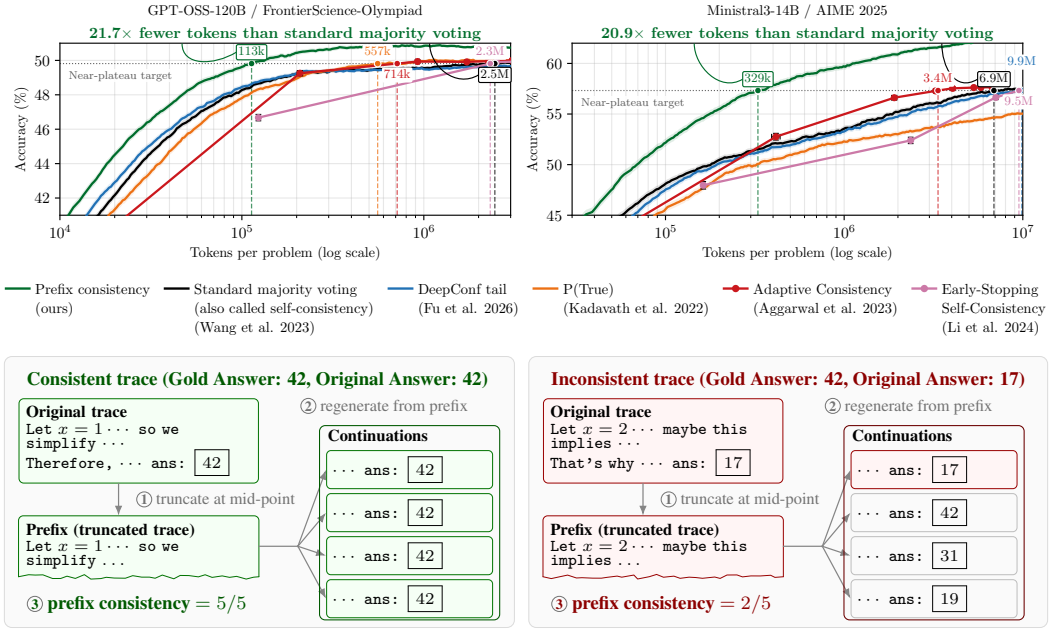


Figure 1: **Overview of prefix consistency.** *Top*: cost-equivalent accuracy of majority voting weighted by prefix consistency and baseline methods on two models and benchmark settings. Shaded bands and error bars show $\pm 2\times$ the Monte Carlo standard error. *Bottom*: Overview of our proposed method (prefix consistency). Correct reasoning traces exhibit greater reproducibility under regeneration.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Large language models (LLMs) have shown strong reasoning ability when allowed to produce chain-of-thought (CoT) reasoning [Kojima et al., 2022, Wei et al., 2022]. Generating intermediate reasoning steps substantially improves performance on challenging tasks such as math [Zhou et al., 2023, Fu et al., 2023], scientific reasoning [Lu et al., 2022, Wang et al., 2024], and knowledge-intensive question answering [Trivedi et al., 2023, Wang et al., 2023a].

A simple and effective way to further improve the accuracy of the final answer is majority voting (MV, also known as self-consistency), which samples a diverse set of CoTs and returns the most frequent answer [Wang et al., 2023b]. A limitation is that Standard MV treats all CoT outputs equally and still fails when the correct answer is in the minority.

To improve MV, the standard approach has been to use weighted majority voting (WMV). WMV refines MV by weighting each generation according to its quality. The more reliable a generation is, the greater the signal it receives. Existing WMV methods derive a per-sample reliability signal from the generated trace, including response probability [Wang et al., 2023b], self-certainty [Kang et al., 2025], DeepConf [Fu et al., 2026], verbalized confidence elicited in text [Lin et al., 2022, Taubenfeld et al., 2025], and P(True) [Kadavath et al., 2022]. Previous studies have demonstrated that these reliability-aware aggregation methods outperform Standard MV. However, these signals often fail to separate correct from wrong traces on difficult problems, the regime where Standard MV most needs improvement (Figure 2).

We introduce a novel reliability signal, **prefix consistency**, and incorporate it into WMV. This signal is motivated by the observation that correct reasoning traces tend to be more reproducible under regeneration than wrong ones. We truncate each sample's CoT at a specified fraction and regenerate continuations from the prefix (Figure 1). Prefix consistency requires no access to token log-probabilities. Since regenerated answers also participate in voting, our method recovers correct answers absent from the initial samples.

Our contributions are:

1. We propose *prefix consistency*, a reliability signal that truncates each sample's CoT and regenerates from the prefix, and use it to form prefix-consistency-weighted majority voting (PC-WMV). PC-WMV requires no access to token log-probabilities.
2. Across 4 benchmarks and 5 model scales, prefix consistency outperforms existing WMV baselines (e.g. DeepConf, P(True), Self-certainty) as a correctness predictor (best macro-averaged AUROC on 15 out of 20 (model, benchmark) cells, .63–.80). On many problems with Pass@1 below 50%, where Standard MV fails by default, prefix consistency still discriminates correct from wrong traces (Section 4.4), leaving room for PC-WMV to find the correct answer.
3. In cost-equivalent comparison against the primary WMV baselines (DeepConf tail, P(True), Self-certainty) and adaptive-stopping baselines (AC, ESC), PC-WMV is the most cost-efficient on the majority of the 20 (model, benchmark) settings, reaching Standard MV plateau at a median $4.6\times$ fewer tokens (up to $21\times$ vs. Standard MV and $10\times$ vs. AC sweep, Figure 1 and Table 4).

2 Preliminary

We consider a benchmark $\mathcal{Q}$, a set of problems. For each problem $q \in \mathcal{Q}$, we have an answer space $\mathcal{A}$ and a correct answer $a_q^\star \in \mathcal{A}$. Given q, an LLM generates a trace y, i.e., a sequence of tokens that represents a CoT followed by a final summary, from which we parse the final answer $a \in \mathcal{A}$. We write $(y, a) \sim \mathrm{LLM}(\cdot \mid q)$. We write $\mathrm{Pass@1}_q := \Pr_{(y,a)\sim\mathrm{LLM}(\cdot|q)}[a = a_q^\star]$ for the per-problem single-sample success probability, and report the macro-average over $\mathcal{Q}$ as the benchmark-level Pass@1. When the context is clear, we suppress q in the notation (e.g., $a^\star$ instead of $a_q^\star$).

To improve the accuracy over Pass@1 at test time, we draw N independent samples $\{(y_i, a_i)\}_{i=1}^N$ and aggregate them into a single output. A standard method that aggregates the N answers is *majority voting* (MV, also known as self-consistency), which returns the most frequent answer:

$$\hat{a}_N^{\mathrm{MV}} = \arg\max_{a\in\mathcal{A}} \sum_{i=1}^{N} \mathbf{1}[a_i = a]. \tag{1}$$

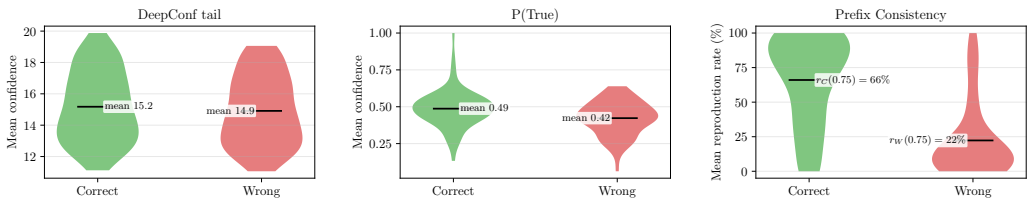


Figure 2: **Per-problem reliability signal distributions for correct vs. wrong traces** on FrontierScience-Olympiad with GPT-OSS-120B. Each violin shows the distribution of per-problem mean signal values across the 73 problems that have at least one correct and one wrong trace. On this difficult benchmark (Pass@1 $\approx$.34), the baseline methods DeepConf tail and P(True) place correct and wrong traces in nearly overlapping ranges, while prefix consistency separates them. The same baselines have some discriminative ability on easier (model, benchmark) settings (Table 2). The point is that they fail precisely in the regime where reweighting matters most.

We refer to this unweighted aggregator as *Standard MV* (Eq. (1)). Standard MV treats all samples equally and fails when the correct answer is not the mode of the answer distribution, typically observed when an LLM faces challenging problems where Pass@1 accuracy is below 50%. A natural extension is *weighted majority voting* (WMV), where each sample i contributes a weighted vote $v_i(a) \geq 0$ for answer a:

$$\hat{a}_N^{\mathrm{WMV}} = \arg\max_{a \in \mathcal{A}} \sum_{i=1}^{N} v_i(a). \tag{2}$$

For sample i, let ℓ_i denote the model's token-level log-probabilities available to the signal. Prior WMV methods extract a confidence signal $s(y_i, \ell_i) \geq 0$ from the trace and apply a weighting function $w \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to it:

$$v_i(a) = w(s(y_i, \ell_i)) \cdot \mathbf{1}[a_i = a]. \tag{3}$$

Another class of WMV methods adopts verbalized signals that require no log-probability access ($\ell_i = \emptyset$), where s depends only on the text of y_i. Other methods (Self-certainty, DeepConf, Response probability) set ℓ_i to the per-token log-probabilities along y_i. P(True) sets ℓ_i to the log-probability of the "True" token under a self-rating prompt. Appendix E gives the explicit form of s and w for each baseline. However, such confidence signals often fail to separate correct traces from wrong traces on difficult problems. We next introduce prefix consistency, a signal that requires no log-probability access ($\ell_i = \emptyset$).

3 Prefix Consistency

We propose prefix consistency, a reliability signal: truncate each sample's CoT at an intermediate point and regenerate its continuation, treating samples whose initial answer reappears as more reliable than those whose answer changes (Figure 1).

3.1 Prefix Consistency as a Reliability Signal

For each sample (y_i, a_i), let $|y_i|$ denote the number of tokens in y_i. We truncate y_i after its first $\lceil \tau |y_i| \rceil$ tokens for a fixed fraction $\tau \in (0, 1)$, and regenerate K continuations from this prefix, yielding the multiset:

$$A_i^{(\tau,K)} = \{a_i, \tilde{a}_{i,1}^{(\tau)}, \ldots, \tilde{a}_{i,K}^{(\tau)}\}. \tag{4}$$

We refer to $A_i^{(\tau,K)}$ as the i-th *group*. For the following discussion, we focus on the case when $K = 1$. We will write $A_i^{(\tau)}$ for $A_i^{(\tau,1)}$ and write $\tilde{a}_i^{(\tau)}$ for $\tilde{a}_{i,1}^{(\tau)}$. Extending this to arbitrary K is straightforward.

The key empirical phenomenon is a *reproduction-rate asymmetry*: a regenerated answer is more likely to match the initial answer when the initial answer is correct. Let $r_C(\tau)$ and $r_W(\tau)$ denote the probabilities of reproducing the initial answer, conditioned on whether it is correct or wrong:

$$r_C(\tau) := \Pr[\tilde{a}_i^{(\tau)} = a_i \mid a_i = a^\star], \qquad r_W(\tau) := \Pr[\tilde{a}_i^{(\tau)} = a_i \mid a_i \neq a^\star] \tag{5}$$

Algorithm 1 Prefix-Consistency-Weighted Majority Voting (PC-WMV)

Require: Problem q, #groups N, truncation fraction τ, #regenerations per group K, weighting w

votes $\leftarrow \emptyset$ {map from answer to accumulated votes}
for $i = 1$ to N **do**
  $y_i \leftarrow \textsc{GenerateTrace}(q)$
  $a_i \leftarrow \textsc{ExtractAnswer}(y_i)$
  $y_i[:\lceil \tau |y_i| \rceil] \leftarrow \textsc{Truncate}(y_i, \tau)$ {shared prefix}
  $A_i^{(\tau,K)} \leftarrow [a_i]$ {multiset of answers for i-th group}
  for $k = 1$ to K **do**
    $\tilde{y}_{i,k}^{(\tau)} \leftarrow \textsc{ContinueFromPrefix}(q, y_i[:\lceil \tau |y_i| \rceil])$
    $\tilde{a}_{i,k}^{(\tau)} \leftarrow \textsc{ExtractAnswer}(\tilde{y}_{i,k}^{(\tau)})$
    append $\tilde{a}_{i,k}^{(\tau)}$ to $A_i^{(\tau,K)}$
  end for
  for each distinct a in $A_i^{(\tau,K)}$ **do**
    $c_i^{(\tau,K)}(a) \leftarrow |\{a' \in A_i^{(\tau,K)} : a' = a\}|/(K{+}1)$ {general-K form of Eq. (7)}
    votes$[a] \leftarrow$ votes$[a] + w(c_i^{(\tau,K)}(a))$
  end for
end for
return $\arg\max_{a \in \mathcal{A}}$ votes$[a]$

Across models and benchmarks, we consistently observe the following inequality (Table 1):

$$r_C(\tau) > r_W(\tau). \tag{6}$$

In other words, a correct initial answer is more likely to be reproduced under regeneration than a wrong initial answer. Figure 2 illustrates this asymmetry on FrontierScience-Olympiad with GPT-OSS-120B and contrasts it with two baseline signals (DeepConf tail and P(True)) that fail to distinguish between correct and wrong traces.

To exploit this observation, we score each candidate $a \in A_i^{(\tau)}$ by its reproducibility within the group:

$$c_i^{(\tau)}(a) := \frac{|\{a' \in A_i^{(\tau)} : a' = a\}|}{2}. \tag{7}$$

We denote the *prefix consistency score* of a in group i by $c_i^{(\tau)}(a) \in \{0, 1/2, 1\}$. Unlike conventional per-sample reliability signals (e.g., DeepConf tail, P(True)), which assign a single scalar to each sample's initial answer, $c_i^{(\tau)}(a)$ is defined for every candidate $a \in A_i^{(\tau)}$, including regenerated answers that did not appear among the initial answers.

3.2 Prefix-Consistency-Weighted Majority Voting (Algorithm 1)

We set the WMV weight in Eq. (2) using Eq. (7):

$$v_i(a) \;=\; w\Big(c_i^{(\tau)}(a)\Big) \tag{8}$$

where $w\colon [0,1] \to \mathbb{R}_{\geq 0}$ with $w(0) = 0$.

We refer to this method as **prefix-consistency-weighted majority voting (PC-WMV)** (Algorithm 1). Since Eq. (8) weights every distinct $a \in A_i^{(\tau)}$ rather than only the initial answer a_i, PC-WMV's aggregated vote $\sum_i v_i(a)$ can be positive for regenerated answers absent from the initial N samples. This is the operational consequence of using a per-candidate signal rather than a per-sample one.

We now demonstrate that this additional flexibility results in a clear advantage over Standard MV in situations where Standard MV is proven to fail. Standard MV fails when the correct answer occurs less frequently than a wrong answer. The following theorem, in the simplest case of a binary answer space, demonstrates how PC-WMV uses the reproduction-rate asymmetry to recover the correct answer in this regime.

Theorem 1 (PC-WMV strictly improves over Standard MV when $r_C(\tau) > r_W(\tau)$)**.** *Fix $\tau \in (0,1)$ and $K = 1$. Let $\mathcal{A} = \{a^\star, a'\}$, where $a^\star$ is the correct answer and a' is the only wrong answer. Assume the N groups are i.i.d., and let $\pi(a^\star) := \Pr[a_i = a^\star]$ denote the Pass@1. Assume $r_C(\tau) > r_W(\tau) > 0$. For any weighting function w with $w(0) = 0$ and $w(1) > 0$, in the limit of $N \to \infty$, PC-WMV converges to $a^\star$ a.s. iff*

$$\pi(a^\star) > \frac{r_W(\tau)}{r_C(\tau) + r_W(\tau)}. \tag{9}$$

In contrast, Standard MV converges to $a^\star$ a.s. if and only if $\pi(a^\star) > \frac{1}{2}$. Therefore, PC-WMV converges to $a^\star$ on the interval where Standard MV does not:

$$\pi(a^\star) \in \left(\frac{r_W(\tau)}{r_C(\tau) + r_W(\tau)}, \frac{1}{2} \right], \tag{10}$$

of width

$$\frac{D(\tau)}{2\left(r_C(\tau) + r_W(\tau)\right)} \qquad (D(\tau) := r_C(\tau) - r_W(\tau)). \tag{11}$$

The formal proof of Theorem 1 is in Appendix C.3.

Interpretation. The key quantity in our analysis is $D(\tau)$, which we call the *discrimination gap*. Theorem 1 considers the simple case of two answer candidates, correct and wrong: when $D(\tau) > 0$, PC-WMV recovers the correct answer when Standard MV does not, provided that Pass@1 falls within the interval of Eq. (10). At fixed $r_C(\tau) + r_W(\tau)$, a larger $D(\tau)$ widens the region in which PC-WMV outperforms Standard MV. Our results (Table 1) show that $D(\tau)$ is substantially larger than 0 across models and benchmarks, so PC-WMV is effective in practice.

Hyperparameters. Prefix consistency has two hyperparameters: the truncation fraction $\tau \in (0,1)$ and the number of regenerations per group $K \in \mathbb{N}$. PC-WMV adds a third, the weighting function $w\colon [0,1] \to \mathbb{R}_{\geq 0}$ with $w(0) = 0$.

4 Experiments

We conduct experiments on science (FrontierScience-Olympiad [Wang et al., 2025]) and math (HMMT Feb 2026, AIME 2025, Brumo 2025 [Balunović et al., 2025]) datasets. We evaluate on five reasoning LLMs: GPT-OSS-120B, GPT-OSS-20B [OpenAI, 2025], Nemotron3-30B (`NVIDIA-Nemotron-3-Nano-30B-A3B-BF16`) [NVIDIA, 2025a], Nemotron2-9B (`NVIDIA-Nemotron-Nano-9B-v2`) [NVIDIA, 2025b], and Ministral3-14B (`Ministral-3-14B-Reasoning-2512`) [Mistral AI, 2026].

We compare our proposed methods against Standard MV, three primary WMV baselines (Self-certainty [Kang et al., 2025], DeepConf tail [Fu et al., 2026], and P(True) [Kadavath et al., 2022, Taubenfeld et al., 2025]), and two adaptive-stopping rules over MV (Adaptive Consistency, AC [Aggarwal et al., 2023], and Early-Stopping Self-Consistency, ESC [Li et al., 2024]). The details of these methods are documented in Appendix E.

Hyperparameters of prefix consistency. Unless otherwise specified, we fix the truncation fraction at $\tau = 0.75$. The results in the main paper use $K = 1$ throughout. We accordingly suppress τ in the notation and write $c_i(a), r_C, r_W, D$ for $c_i^{(\tau)}(a), r_C(\tau), r_W(\tau), D(\tau)$.

4.1 Prefix Consistency as a Correctness Predictor

We report $\overline{\text{AUROC}}$, the macro-averaged AUROC over problems with at least one correct and one wrong initial sample (formal definition in Appendix G.2).

Note that some previous work [Xiong et al., 2024, Fadeeva et al., 2024] adopted AUROC pooled across problems, which differs from $\overline{\text{AUROC}}$. However, as argued in Taubenfeld et al. [2025], such an

Table 1: **Reproduction rates r_C, r_W and discrimination gap $D = r_C - r_W$ for prefix consistency (larger D is better).** Macro-averaged over problems with at least one correct and one wrong initial sample. $r_C \geq r_W$ holds on every (model, benchmark) cell, and a larger D predicts a larger PC-WMV advantage over Standard MV (Theorem 1).

	GPT-OSS-120B			GPT-OSS-20B			Nemotron3-30B			Nemotron2-9B			Ministral3-14B		
Benchmark	r_C	r_W	D	r_C	r_W	D	r_C	r_W	D	r_C	r_W	D	r_C	r_W	D
FrontierScience-Olympiad	66.0	22.3	43.7	55.4	14.0	41.4	50.9	24.7	26.2	45.0	37.5	7.4	56.4	26.0	30.4
HMMT Feb 2026	87.8	48.1	39.7	78.3	29.7	48.6	80.0	40.4	39.6	79.3	49.6	29.7	74.5	32.0	42.4
AIME 2025	95.5	61.7	33.7	87.8	42.6	45.2	74.7	38.2	36.5	69.5	43.2	26.3	77.3	26.1	51.1
Brumo 2025	83.8	56.7	27.1	82.8	39.0	43.7	75.6	15.3	60.3	81.2	54.7	26.6	76.2	24.4	51.8

Table 2: $\overline{\text{AUROC}}$ **for correctness discrimination (higher is better).** Macro-averaged $\overline{\text{AUROC}}$ per (model, benchmark), with the best per column in bold.

	GPT-OSS-120B				GPT-OSS-20B				Nemotron3-30B				Nemotron2-9B				Ministral3-14B			
Signal	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo
Prefix consistency	**.719**	.698	.669	**.636**	**.707**	**.743**	**.726**	**.719**	**.631**	**.698**	**.682**	**.801**	.537	**.648**	.631	.633	**.652**	**.712**	**.756**	**.759**
Self-certainty	.570	.664	.641	.549	.545	.570	.560	.484	.496	.414	.520	.655	.458	.605	.537	.580	.318	.359	.408	.342
DeepConf bottom-10%	.525	.640	.601	.486	.525	.580	.565	.481	.474	.413	.514	.607	.434	.515	.529	.515	.342	.331	.395	.347
DeepConf tail	.565	**.728**	**.703**	.606	.583	.659	.645	.557	.508	.588	.465	.755	.494	.628	.665	.641	.349	.404	.421	.358
Verbal 0–100	.516	.552	.505	.436	.516	.574	.561	.511	.524	.489	.606	.532	.571	.563	.584	.548	.505	.525	.525	.450
P(True)	.568	.550	.551	.527	.587	.667	.592	.583	.493	.532	.578	.564	**.620**	.635	**.724**	**.708**	.469	.454	.441	.477

Abbreviations: FSci = FrontierScience-Olympiad, HMMT = HMMT Feb 2026, AIME = AIME 2025, Brumo = Brumo 2025.

AUROC pooled across problems conflates within-problem discrimination with cross-problem score-difficulty correlation, and only the former predicts whether confidence-weighted self-consistency improves over Standard MV. They also report that calibration metrics such as Expected Calibration Error and Brier score are similarly unsuitable. In their data, the best-calibrated source (verbalized binary) gave the smallest improvement while the strongest method (P(True) therein) was only moderately calibrated. At vote time, every WMV method (including PC-WMV and all baselines) compares scores only among samples from the same problem, and thus we consider $\overline{\text{AUROC}}$ to better measure discriminative ability between correct traces and wrong traces.

Table 1 reports the discrimination gap $D = r_C - r_W$ for prefix consistency: $D > 0$ on every (model, benchmark) cell, confirming the asymmetry $r_C > r_W$. Table 2 reports $\overline{\text{AUROC}}$ for prefix consistency against the WMV baselines. Prefix consistency has the highest $\overline{\text{AUROC}}$ on 15 of 20 cells (typically around 0.7), separating correct from wrong traces more clearly than the baselines. Baselines' $\overline{\text{AUROC}}$ often hovers near 0.5 (= random[1]) on harder cells, where their scores differ little between correct and wrong samples, reaching ~ 0.7 only on some easier cells.

4.2 Weighted Majority Voting Results

We compare PC-WMV against existing WMV methods under the same generated-token budget.

Weighting variants. We use the power family $w^{(n)}(c) = c^n$ for $n \in \{1, 2, 3\}$, denoted PC-linear, PC-quadratic, and PC-cubic, where the "PC" prefix abbreviates prefix consistency. Under $K = 1$, $c_i(a) \in \{0, 1/2, 1\}$. Thus, a candidate that is reproduced under regeneration ($c = 1$) receives weight 1, while a candidate that appears in only one of the two traces ($c = 1/2$) receives weight $1/2^n$. The weight ratio between a reproduced candidate and a single-trace one is therefore $2^n : 1$, that is, $2 : 1$ for linear, $4 : 1$ for quadratic, and $8 : 1$ for cubic. The larger n is, the more pronounced the relative weight given to reproduced answers.

Cost-equivalent evaluation. We measure inference cost by the total number of generated tokens, treating each generated token as equally expensive, and log-probability access as free. For each model and benchmark, we first generate N=128 initial samples per problem (N=64 for Ministral3-14B), from which all methods sample with replacement under a common token budget B. See Appendix G for the pool construction, trial design, and confidence-interval definition.

[1] $\overline{\text{AUROC}} = 0.5$ is the value attained by a random signal that is uninformative about correctness.

Table 3: **Weighted majority voting accuracy at fixed token budget B (higher is better).** Each method's accuracy at $B \in \{250\text{k}, 1\text{M}, 5\text{M}\}$ tokens, with the best per (model, B) column in bold.

		GPT-OSS-120B			Nemotron3-30B			Ministral3-14B		
Benchmark	Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
FrontierScience-Olympiad	Standard MV	$.493_{\pm.001}$	$.495_{\pm.001}$	$.499_{\pm.001}$	$.404_{\pm.002}$	$.468_{\pm.002}$	$.490_{\pm.001}$	$.137_{\pm.001}$	$.141_{\pm.001}$	$.142_{\pm.001}$
	Self-certainty	$.494_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.402_{\pm.002}$	$.468_{\pm.002}$	$.491_{\pm.001}$	$.130_{\pm.001}$	$.135_{\pm.001}$	$.137_{\pm.001}$
	DeepConf tail	$.493_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.405_{\pm.002}$	$.470_{\pm.002}$	$.494_{\pm.001}$	$.130_{\pm.001}$	$.134_{\pm.001}$	$.137_{\pm.001}$
	P(True)	$.493_{\pm.001}$	$.500_{\pm.001}$	$.500_{\pm.000}$	$.394_{\pm.003}$	$.465_{\pm.002}$	$.492_{\pm.002}$	$.137_{\pm.001}$	$.143_{\pm.001}$	$.146_{\pm.001}$
	PC-linear	$.495_{\pm.001}$	$.496_{\pm.001}$	$.496_{\pm.001}$	$.429_{\pm.002}$	$.481_{\pm.002}$	$.494_{\pm.001}$	$.149_{\pm.002}$	$.160_{\pm.001}$	$.162_{\pm.001}$
	PC-quadratic	$.502_{\pm.001}$	$.504_{\pm.001}$	$.503_{\pm.001}$	$\mathbf{.433}_{\pm.002}$	$\mathbf{.486}_{\pm.002}$	$\mathbf{.504}_{\pm.002}$	$.156_{\pm.002}$	$.170_{\pm.001}$	$.173_{\pm.001}$
	PC-cubic	$\mathbf{.506}_{\pm.001}$	$\mathbf{.508}_{\pm.001}$	$\mathbf{.507}_{\pm.001}$	$\mathbf{.433}_{\pm.002}$	$\mathbf{.486}_{\pm.002}$	$.503_{\pm.001}$	$\mathbf{.158}_{\pm.002}$	$\mathbf{.174}_{\pm.001}$	$\mathbf{.181}_{\pm.001}$
HMMT Feb 2026	Standard MV	$.708_{\pm.003}$	$.745_{\pm.003}$	$.763_{\pm.001}$	$.736_{\pm.004}$	$.777_{\pm.003}$	$.802_{\pm.002}$	$.413_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.458_{\pm.002}$
	Self-certainty	$.713_{\pm.003}$	$.749_{\pm.003}$	$.768_{\pm.002}$	$.734_{\pm.004}$	$.779_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.409_{\pm.004}$	$.429_{\pm.003}$	$.438_{\pm.002}$
	DeepConf tail	$.720_{\pm.003}$	$.758_{\pm.003}$	$.783_{\pm.002}$	$.740_{\pm.004}$	$.779_{\pm.003}$	$.801_{\pm.002}$	$.408_{\pm.004}$	$.427_{\pm.003}$	$.436_{\pm.002}$
	P(True)	$.706_{\pm.003}$	$.743_{\pm.003}$	$.784_{\pm.002}$	$.736_{\pm.004}$	$.779_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.404_{\pm.003}$	$.429_{\pm.003}$	$.433_{\pm.002}$
	PC-linear	$.717_{\pm.003}$	$.751_{\pm.002}$	$.764_{\pm.001}$	$\mathbf{.742}_{\pm.004}$	$.785_{\pm.003}$	$.802_{\pm.002}$	$\mathbf{.418}_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.444_{\pm.002}$
	PC-quadratic	$.725_{\pm.003}$	$.764_{\pm.002}$	$.782_{\pm.001}$	$\mathbf{.742}_{\pm.004}$	$\mathbf{.787}_{\pm.003}$	$.803_{\pm.002}$	$.415_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.447_{\pm.002}$
	PC-cubic	$\mathbf{.727}_{\pm.003}$	$\mathbf{.771}_{\pm.002}$	$\mathbf{.786}_{\pm.001}$	$\mathbf{.742}_{\pm.004}$	$\mathbf{.787}_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.414_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$\mathbf{.461}_{\pm.002}$
AIME 2025	Standard MV	$.901_{\pm.002}$	$.901_{\pm.001}$	$.900_{\pm.000}$	$.935_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.508_{\pm.003}$	$.535_{\pm.003}$	$.569_{\pm.002}$
	Self-certainty	$.904_{\pm.002}$	$.903_{\pm.001}$	$.900_{\pm.000}$	$.935_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.505_{\pm.003}$	$.533_{\pm.003}$	$.565_{\pm.002}$
	DeepConf tail	$.906_{\pm.002}$	$.905_{\pm.001}$	$.900_{\pm.000}$	$\mathbf{.937}_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.506_{\pm.003}$	$.533_{\pm.003}$	$.565_{\pm.002}$
	P(True)	$.902_{\pm.002}$	$.909_{\pm.001}$	$.906_{\pm.001}$	$.930_{\pm.003}$	$.964_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.495_{\pm.003}$	$.523_{\pm.003}$	$.542_{\pm.002}$
	PC-linear	$.906_{\pm.002}$	$.913_{\pm.002}$	$.909_{\pm.001}$	$.932_{\pm.003}$	$.964_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.556_{\pm.003}$	$.583_{\pm.003}$	$.600_{\pm.002}$
	PC-quadratic	$.912_{\pm.002}$	$.922_{\pm.002}$	$.930_{\pm.002}$	$.932_{\pm.003}$	$.963_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$\mathbf{.565}_{\pm.003}$	$.595_{\pm.003}$	$.616_{\pm.002}$
	PC-cubic	$\mathbf{.913}_{\pm.002}$	$\mathbf{.926}_{\pm.002}$	$\mathbf{.941}_{\pm.002}$	$.932_{\pm.003}$	$.963_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.564_{\pm.003}$	$\mathbf{.597}_{\pm.003}$	$\mathbf{.621}_{\pm.002}$
Brumo 2025	Standard MV	$.801_{\pm.003}$	$.821_{\pm.002}$	$.833_{\pm.000}$	$.896_{\pm.003}$	$.928_{\pm.002}$	$.931_{\pm.001}$	$.669_{\pm.004}$	$.691_{\pm.003}$	$.699_{\pm.003}$
	Self-certainty	$.805_{\pm.003}$	$.826_{\pm.002}$	$.833_{\pm.000}$	$.918_{\pm.003}$	$.954_{\pm.002}$	$.963_{\pm.001}$	$.659_{\pm.004}$	$.682_{\pm.003}$	$.686_{\pm.002}$
	DeepConf tail	$.810_{\pm.003}$	$.831_{\pm.001}$	$.833_{\pm.000}$	$.920_{\pm.003}$	$.955_{\pm.002}$	$.965_{\pm.001}$	$.659_{\pm.004}$	$.681_{\pm.003}$	$.683_{\pm.002}$
	P(True)	$.794_{\pm.003}$	$.821_{\pm.003}$	$.844_{\pm.002}$	$.886_{\pm.003}$	$.917_{\pm.002}$	$.929_{\pm.001}$	$.632_{\pm.004}$	$.658_{\pm.003}$	$.661_{\pm.003}$
	PC-linear	$.810_{\pm.003}$	$.830_{\pm.002}$	$.833_{\pm.000}$	$\mathbf{.933}_{\pm.003}$	$\mathbf{.961}_{\pm.002}$	$\mathbf{.968}_{\pm.001}$	$.675_{\pm.004}$	$.705_{\pm.003}$	$.723_{\pm.002}$
	PC-quadratic	$.815_{\pm.003}$	$.837_{\pm.002}$	$.842_{\pm.001}$	$.932_{\pm.003}$	$.959_{\pm.002}$	$\mathbf{.968}_{\pm.001}$	$.684_{\pm.004}$	$.719_{\pm.003}$	$.740_{\pm.002}$
	PC-cubic	$\mathbf{.818}_{\pm.003}$	$\mathbf{.842}_{\pm.002}$	$\mathbf{.857}_{\pm.001}$	$.929_{\pm.003}$	$.953_{\pm.002}$	$.964_{\pm.001}$	$\mathbf{.686}_{\pm.004}$	$\mathbf{.727}_{\pm.003}$	$\mathbf{.759}_{\pm.003}$

Table 3 reports accuracy under fixed token budgets (250k, 1M, and 5M tokens) across the three models and four benchmarks. At 1M tokens, prefix consistency matches or exceeds all baselines on the more difficult benchmarks (FrontierScience-Olympiad, HMMT Feb 2026), while the advantage is smaller on AIME 2025 where Standard MV already achieves high accuracy. The improvements are consistent across weighting functions (PC-linear, PC-quadratic, and PC-cubic). PC-cubic provides the greatest improvement for the most difficult problems. Per-model tables with the full set of baselines (DeepConf variants, Response probability, verbalized confidence, P(True)) for all five models, including the two not shown above (GPT-OSS-20B, Nemotron2-9B), are reported in Appendix D.3.

4.3 Token Efficiency

Section 4.2 reported accuracy under fixed budget constraints. We next compare how many tokens each method needs to reach the same target accuracy.

Table 4 shows the token-efficiency ratio $B_{\text{method}}/B_{\text{MV}}$, where B_X is the budget method X needs to reach the target accuracy Pass@1 $+ \alpha\times$ (Standard MV plateau $-$ Pass@1) for $\alpha \in \{75\%, 90\%, 99\%\}$. The Standard MV plateau is Standard MV's bootstrap-saturated accuracy on the N-sample pool (Appendix G.3), so α interpolates between Pass@1 (α=0) and this plateau (α=1). A ratio < 1 indicates a cost advantage for X over Standard MV at the target. For example, PC-cubic, shown as "Prefix consistency" in Figure 1, achieves a token-efficiency ratio of approximately 0.05, corresponding to a $21\times$ reduction in tokens. The headline numbers in this paper (median $4.6\times$, up to $21\times$ vs. Standard MV, up to $10\times$ vs. AC sweep) are computed at α=99% across all 20 (model, benchmark) cells: Table 4 together with Table 12 (GPT-OSS-20B) and Table 14 (Nemotron2-9B) in Appendix D.3.

PC-cubic is more cost-efficient than Standard MV on 11 out of 12 model and benchmark settings at α=99%. The strongest savings are $0.05\times$ on both FrontierScience-Olympiad with GPT-OSS-120B and AIME 2025 with Ministral3-14B, and $0.08\times$ on FrontierScience-Olympiad with Ministral3-14B. All three correspond to cells with large D (Table 1).

The savings track the discrimination gap D (Table 1): pairs with the largest D yield the largest reductions. PC-cubic offers little advantage over Standard MV on two cells (it underperforms on Nemotron3-30B AIME 2025 and only marginally beats Standard MV on Nemotron3-30B HMMT Feb 2026), both of which have a small Pass@1-to-plateau gap of at most .11 (.902 $\rightarrow$ .967 and .708 $\rightarrow$ .810), leaving little room above Pass@1 for any reweighting irrespective of D. The practical penalty in this regime, where Pass@1 is close to Standard MV plateau, is correspondingly small.

Table 4: **Token efficiency ratio $B_{\text{method}}/B_{\text{MV}}$ at target accuracy α between Pass@1 and the Standard MV plateau (smaller is better).** B_X is the budget method X needs to reach Pass@$1 + \alpha\times$ (Standard MV plateau $-$ Pass@1) for $\alpha \in \{75\%, 90\%, 99\%\}$. Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates the method's plateau is below the target. Best method per column in bold.

		GPT-OSS-120B			Nemotron3-30B			Ministral3-14B		
Benchmark	Method	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
FrontierScience-Olympiad	Pass@1 / Standard MV plateau	.338 / .500			.295 / .493			.091 / .142		
	Standard MV budget (B_{MV})	45k	107k	2.5M	486k	1.3M	6.0M	118k	249k	924k
	Self-certainty	$0.89_{\pm 0.04}\times$	$0.87_{\pm 0.06}\times$	N/A	$1.04_{\pm 0.06}\times$	$1.01_{\pm 0.11}\times$	$0.85_{\pm 0.19}\times$	$1.62_{\pm 0.20}\times$	$>10\times$	N/A
	DeepConf tail	$0.91_{\pm 0.05}\times$	$0.91_{\pm 0.07}\times$	N/A	$0.97_{\pm 0.06}\times$	$0.88_{\pm 0.08}\times$	$0.56_{\pm 0.14}\times$	$1.54_{\pm 0.18}\times$	$>10\times$	N/A
	P(True)	$1.23_{\pm 0.05}\times$	$1.20_{\pm 0.10}\times$	$0.23_{\pm 0.09}\times$	$1.14_{\pm 0.06}\times$	$1.10_{\pm 0.11}\times$	$0.75_{\pm 0.18}\times$	$1.10_{\pm 0.12}\times$	$1.01_{\pm 0.27}\times$	$0.48_{\pm 0.30}\times$
	AC sweep	$1.93_{\pm 0.09}\times$	$1.54_{\pm 0.12}\times$	$0.29_{\pm 0.15}\times$	$1.24_{\pm 0.07}\times$	$0.87_{\pm 0.09}\times$	$0.49_{\pm 0.13}\times$	$2.48_{\pm 0.40}\times$	$4.28_{\pm 1.45}\times$	N/A
	ESC sweep	$2.74_{\pm 0.08}\times$	$5.59_{\pm 0.70}\times$	$0.94_{\pm 0.53}\times$	$2.06_{\pm 0.26}\times$	$3.62_{\pm 0.41}\times$	$2.23_{\pm 0.58}\times$	$4.53_{\pm 1.55}\times$	$>10\times$	$>10\times$
	PC-linear	$0.70_{\pm 0.03}\times$	$0.63_{\pm 0.05}\times$	$3.88_{\pm 2.37}\times$	$0.67_{\pm 0.04}\times$	$0.54_{\pm 0.05}\times$	$0.40_{\pm 0.09}\times$	$\mathbf{0.38}_{\pm 0.04}\times$	$0.26_{\pm 0.07}\times$	$0.10_{\pm 0.07}\times$
	PC-quadratic	$\mathbf{0.65}_{\pm 0.03}\times$	$\mathbf{0.52}_{\pm 0.04}\times$	$0.06_{\pm 0.02}\times$	$\mathbf{0.61}_{\pm 0.03}\times$	$\mathbf{0.48}_{\pm 0.04}\times$	$0.24_{\pm 0.06}\times$	$\mathbf{0.38}_{\pm 0.04}\times$	$\mathbf{0.24}_{\pm 0.06}\times$	$\mathbf{0.08}_{\pm 0.05}\times$
	PC-cubic	$\mathbf{0.65}_{\pm 0.03}\times$	$\mathbf{0.52}_{\pm 0.04}\times$	$\mathbf{0.05}_{\pm 0.02}\times$	$\mathbf{0.61}_{\pm 0.03}\times$	$\mathbf{0.48}_{\pm 0.04}\times$	$\mathbf{0.22}_{\pm 0.05}\times$	$\mathbf{0.38}_{\pm 0.04}\times$	$\mathbf{0.24}_{\pm 0.06}\times$	$\mathbf{0.08}_{\pm 0.05}\times$
HMMT Feb 2026	Pass@1 / Standard MV plateau	.589 / .760			.708 / .810			.270 / .463		
	Standard MV budget (B_{MV})	364k	858k	1.8M	1.6M	4.2M	8.5M	258k	1.2M	7.4M
	Self-certainty	$0.74_{\pm 0.07}\times$	$0.85_{\pm 0.07}\times$	$0.87_{\pm 0.11}\times$	$0.85_{\pm 0.15}\times$	$0.89_{\pm 0.14}\times$	$0.93_{\pm 0.12}\times$	$1.29_{\pm 0.17}\times$	N/A	N/A
	DeepConf tail	$0.63_{\pm 0.09}\times$	$0.72_{\pm 0.08}\times$	$0.57_{\pm 0.07}\times$	$0.82_{\pm 0.15}\times$	$0.97_{\pm 0.19}\times$	N/A	$1.29_{\pm 0.17}\times$	N/A	N/A
	P(True)	$1.05_{\pm 0.14}\times$	$1.17_{\pm 0.14}\times$	$0.94_{\pm 0.09}\times$	$0.86_{\pm 0.15}\times$	$0.90_{\pm 0.12}\times$	$0.82_{\pm 0.10}\times$	$1.33_{\pm 0.14}\times$	N/A	N/A
	AC sweep	$\mathbf{0.42}_{\pm 0.05}\times$	$0.63_{\pm 0.10}\times$	N/A	$\mathbf{0.51}_{\pm 0.09}\times$	$\mathbf{0.34}_{\pm 0.05}\times$	$\mathbf{0.24}_{\pm 0.03}\times$	$1.52_{\pm 0.28}\times$	$1.21_{\pm 0.28}\times$	$\mathbf{0.60}_{\pm 0.20}\times$
	ESC sweep	$1.96_{\pm 0.33}\times$	$6.66_{\pm 1.29}\times$	N/A	$4.33_{\pm 1.24}\times$	$3.12_{\pm 0.46}\times$	$1.88_{\pm 0.26}\times$	$8.79_{\pm 1.33}\times$	$4.75_{\pm 1.03}\times$	$2.05_{\pm 0.75}\times$
	PC-linear	$0.70_{\pm 0.08}\times$	$0.79_{\pm 0.08}\times$	$0.88_{\pm 0.11}\times$	$0.62_{\pm 0.11}\times$	$0.86_{\pm 0.16}\times$	N/A	$\mathbf{0.85}_{\pm 0.11}\times$	$1.11_{\pm 0.28}\times$	N/A
	PC-quadratic	$0.52_{\pm 0.05}\times$	$0.49_{\pm 0.04}\times$	$0.41_{\pm 0.04}\times$	$0.56_{\pm 0.10}\times$	$0.80_{\pm 0.15}\times$	$1.10_{\pm 0.15}\times$	$0.91_{\pm 0.11}\times$	$1.01_{\pm 0.26}\times$	N/A
	PC-cubic	$0.52_{\pm 0.05}\times$	$\mathbf{0.46}_{\pm 0.04}\times$	$\mathbf{0.35}_{\pm 0.04}\times$	$0.59_{\pm 0.10}\times$	$0.79_{\pm 0.13}\times$	$0.95_{\pm 0.14}\times$	$1.03_{\pm 0.12}\times$	$0.98_{\pm 0.20}\times$	$0.67_{\pm 0.20}\times$
AIME 2025	Pass@1 / Standard MV plateau	.789 / .900			.902 / .967			.345 / .576		
	Standard MV budget (B_{MV})	56k	99k	189k	396k	589k	1.2M	371k	2.0M	6.9M
	Self-certainty	$0.81_{\pm 0.06}\times$	$0.81_{\pm 0.07}\times$	$0.66_{\pm 0.11}\times$	$0.97_{\pm 0.06}\times$	$0.96_{\pm 0.06}\times$	$0.84_{\pm 0.12}\times$	$1.21_{\pm 0.15}\times$	$1.20_{\pm 0.12}\times$	N/A
	DeepConf tail	$\mathbf{0.68}_{\pm 0.04}\times$	$\mathbf{0.70}_{\pm 0.06}\times$	$0.56_{\pm 0.09}\times$	$0.95_{\pm 0.06}\times$	$0.95_{\pm 0.06}\times$	$0.89_{\pm 0.13}\times$	$1.22_{\pm 0.15}\times$	$1.20_{\pm 0.12}\times$	$1.43_{\pm 0.26}\times$
	P(True)	$1.14_{\pm 0.08}\times$	$1.13_{\pm 0.09}\times$	$1.01_{\pm 0.16}\times$	$1.12_{\pm 0.07}\times$	$1.15_{\pm 0.07}\times$	$1.20_{\pm 0.20}\times$	$1.93_{\pm 0.22}\times$	N/A	N/A
	AC sweep	$1.03_{\pm 0.12}\times$	$1.06_{\pm 0.26}\times$	$2.98_{\pm 0.82}\times$	$\mathbf{0.21}_{\pm 0.01}\times$	$\mathbf{0.18}_{\pm 0.01}\times$	$\mathbf{0.21}_{\pm 0.05}\times$	$0.91_{\pm 0.11}\times$	$0.54_{\pm 0.06}\times$	$0.49_{\pm 0.10}\times$
	ESC sweep	$0.96_{\pm 0.06}\times$	$1.38_{\pm 0.57}\times$	$>10\times$	$0.28_{\pm 0.01}\times$	$0.24_{\pm 0.05}\times$	$0.59_{\pm 0.18}\times$	$4.44_{\pm 0.83}\times$	$2.41_{\pm 0.22}\times$	$1.38_{\pm 0.26}\times$
	PC-linear	$0.87_{\pm 0.05}\times$	$0.73_{\pm 0.06}\times$	$0.58_{\pm 0.10}\times$	$1.13_{\pm 0.08}\times$	$1.21_{\pm 0.08}\times$	$1.29_{\pm 0.24}\times$	$\mathbf{0.25}_{\pm 0.03}\times$	$0.11_{\pm 0.01}\times$	$0.07_{\pm 0.01}\times$
	PC-quadratic	$0.86_{\pm 0.05}\times$	$\mathbf{0.70}_{\pm 0.06}\times$	$\mathbf{0.53}_{\pm 0.08}\times$	$1.22_{\pm 0.09}\times$	$1.33_{\pm 0.10}\times$	$1.44_{\pm 0.24}\times$	$0.26_{\pm 0.03}\times$	$\mathbf{0.09}_{\pm 0.01}\times$	$\mathbf{0.05}_{\pm 0.01}\times$
	PC-cubic	$0.86_{\pm 0.05}\times$	$0.71_{\pm 0.06}\times$	$\mathbf{0.53}_{\pm 0.08}\times$	$1.23_{\pm 0.08}\times$	$1.39_{\pm 0.10}\times$	$1.67_{\pm 0.26}\times$	$0.27_{\pm 0.03}\times$	$\mathbf{0.09}_{\pm 0.01}\times$	$\mathbf{0.05}_{\pm 0.01}\times$
Brumo 2025	Pass@1 / Standard MV plateau	.713 / .833			.816 / .932			.421 / .698		
	Standard MV budget (B_{MV})	274k	934k	3.3M	342k	651k	2.5M	92k	265k	1.8M
	Self-certainty	$0.73_{\pm 0.18}\times$	$0.76_{\pm 0.10}\times$	$0.59_{\pm 0.09}\times$	$0.45_{\pm 0.05}\times$	$0.41_{\pm 0.05}\times$	$0.14_{\pm 0.07}\times$	$1.35_{\pm 0.11}\times$	$1.47_{\pm 0.22}\times$	N/A
	DeepConf tail	$0.50_{\pm 0.12}\times$	$0.47_{\pm 0.06}\times$	$0.35_{\pm 0.06}\times$	$0.40_{\pm 0.05}\times$	$0.39_{\pm 0.07}\times$	$0.14_{\pm 0.07}\times$	$1.35_{\pm 0.11}\times$	$1.56_{\pm 0.23}\times$	N/A
	P(True)	$1.46_{\pm 0.36}\times$	$1.07_{\pm 0.11}\times$	$0.60_{\pm 0.11}\times$	$1.40_{\pm 0.14}\times$	$2.05_{\pm 0.31}\times$	$3.39_{\pm 1.62}\times$	$2.33_{\pm 0.22}\times$	N/A	N/A
	AC sweep	$\mathbf{0.32}_{\pm 0.09}\times$	$\mathbf{0.26}_{\pm 0.03}\times$	$0.20_{\pm 0.04}\times$	$\mathbf{0.27}_{\pm 0.03}\times$	$\mathbf{0.18}_{\pm 0.03}\times$	$0.12_{\pm 0.07}\times$	$1.42_{\pm 0.14}\times$	$1.85_{\pm 0.32}\times$	$1.20_{\pm 0.75}\times$
	ESC sweep	$3.19_{\pm 1.47}\times$	$6.11_{\pm 0.65}\times$	N/A	$0.30_{\pm 0.03}\times$	$0.34_{\pm 0.06}\times$	$0.25_{\pm 0.21}\times$	$2.39_{\pm 0.38}\times$	$7.42_{\pm 2.14}\times$	$7.95_{\pm 4.31}\times$
	PC-linear	$0.56_{\pm 0.15}\times$	$0.51_{\pm 0.07}\times$	$0.39_{\pm 0.06}\times$	$0.34_{\pm 0.04}\times$	$0.28_{\pm 0.04}\times$	$\mathbf{0.09}_{\pm 0.05}\times$	$0.93_{\pm 0.07}\times$	$0.80_{\pm 0.12}\times$	$0.32_{\pm 0.19}\times$
	PC-quadratic	$0.44_{\pm 0.11}\times$	$0.37_{\pm 0.04}\times$	$0.19_{\pm 0.03}\times$	$0.33_{\pm 0.03}\times$	$0.28_{\pm 0.04}\times$	$0.10_{\pm 0.05}\times$	$\mathbf{0.86}_{\pm 0.06}\times$	$\mathbf{0.61}_{\pm 0.09}\times$	$0.20_{\pm 0.12}\times$
	PC-cubic	$0.44_{\pm 0.11}\times$	$0.32_{\pm 0.04}\times$	$\mathbf{0.16}_{\pm 0.02}\times$	$0.35_{\pm 0.04}\times$	$0.30_{\pm 0.04}\times$	$0.11_{\pm 0.06}\times$	$\mathbf{0.86}_{\pm 0.06}\times$	$0.62_{\pm 0.09}\times$	$\mathbf{0.18}_{\pm 0.11}\times$

Comparison with adaptive-stopping baselines. PC-cubic is competitive with or better than AC sweep on most cells and outperforms ESC sweep at α=99% on every (model, benchmark) cell except Nemotron3-30B on AIME 2025, despite being a non-adaptive reweighting of the same initial pool that AC and ESC consume sequentially. The advantage is most pronounced on large-D cells (FrontierScience-Olympiad on every model, and most Ministral3-14B benchmarks), e.g. PC-cubic at $0.05\times$ vs. AC sweep at $0.29\times$ at α=99% on GPT-OSS-120B FrontierScience-Olympiad.

AC substantially outperforms PC only on the two cells noted above where Pass@1 is close to Standard MV plateau (Nemotron3-30B on AIME 2025 and HMMT Feb 2026), since the aggregated vote on wrong answers is small and AC's early stop alone bounds the cost. However, the evaluated AC sweep does not reach the α=99% target on 2 out of 12 (model, benchmark) cells (marked "N/A" in Table 4). PC-cubic reaches the target on all 12.

ESC stops at the first fixed-size window of samples that all share the same answer, which is too strict on benchmarks where wrong answers are diverse, so ESC either stops well after AC or fails to stop within the budget. PC and adaptive stopping act on orthogonal axes: PC reweights votes while AC and ESC decide when to stop sampling. A hybrid that votes by PC weights and stops by the AC rule would combine AC's cost bound on easy cells with PC's accuracy on difficult ones (left to future work).

Remark on the cost accounting. Treating log-probability access as free is implementation-dependent but holds for our vLLM setup. This favors baselines that read log-probabilities of the initial trace without generating extra tokens (Self-certainty, DeepConf, Response probability), while prefix consistency spends budget on the regeneration tokens. Even so, the PC-WMV has significant advantage. Imposing any cost for log-probability retrieval only widens the margin.

Additional results on token-efficiency ratios are reported per model in Appendix D.3.

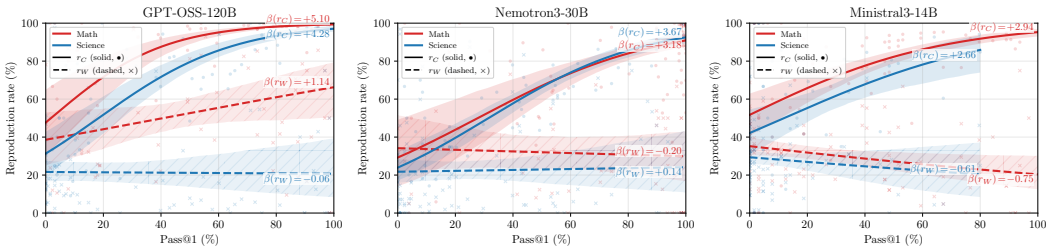


Figure 3: **Per-problem reproduction rates vs. Pass@1.** One panel per model. Solid (•): r_C. Dashed (×): r_W. Both are per-category logistic-regression fits with shaded pointwise cluster-bootstrap confidence intervals (CIs) over problems and per-problem scatter overlays; curves are labeled with the GLM slope β. $r_C > r_W$ holds across the full Pass@1 range, including below 50%. Science: FrontierScience-Olympiad. Math: HMMT Feb 2026 ∪ AIME 2025 ∪ Brumo 2025.

4.4 How Problem Difficulty Affects the Discrimination Gap D

The discrimination gap D determines where prefix consistency improves WMV (Theorem 1), so we now study how D varies with problem difficulty, indexed by Pass@1.

Figure 3 plots per-problem r_C and r_W as a function of Pass@1 for three of the five models (the remaining two are in Appendix D.7.1), stratified by category. Across all five models, r_C rises with problem easiness, r_W depends on the model and category but only weakly on Pass@1, and $D = r_C - r_W$ inherits both. We discuss r_C and r_W in turn below.

First, r_C (solid lines) increases with Pass@1 for both categories. Pass@1 here indexes problem easiness within a fixed model, and on easier problems, the correct answer is more reliably reproduced under regeneration. Logistic generalized linear model (GLM) slopes $\beta(r_C)$ on $\text{logit}(r) = \beta_0 + \beta \cdot \text{Pass@1}$ range from $+2.7$ to $+5.1$ across the six (model, category) pairs shown in Figure 3, all significantly positive (cluster-bootstrap $p < 0.005$, see Appendix D.7.1).

Second, r_W (dashed lines) depends more weakly on Pass@1 than r_C: $|\beta(r_W)| \leq 1.14$ across the six pairs, and among four out of six curves, we cannot statistically reject $\beta = 0$ at a 2σ-equivalent confidence level. For the remaining two non-zero slopes, β is $+1.14$ (GPT-OSS-120B Math) and -0.75 (Ministral3-14B Math), with opposite signs, both smaller in magnitude than the smallest r_C slope. The level of r_W varies by category, sitting at ∼15–40% on Science and ∼25–60% on Math. Math errors may reflect internally consistent miscalculations and science errors may reflect more diffuse knowledge gaps, but this is a hypothesis: our results establish $r_C > r_W$ as a behavioral regularity, and a mechanistic verification (e.g., calculation-heavy vs. concept-heavy subsets) is left to future work.

The finding that D widens on easier cells may appear to conflict with the "savings track D" claim of Section 4.3, since easier cells also have a smaller gap between Pass@1 and Standard MV plateau. The two reconcile by noting that PC-WMV's advantage over Standard MV depends on both D and the gap above Pass@1: a large D pays off only when there is room to reweight votes, which is why the two cells where PC-cubic offers little advantage over Standard MV are precisely the cells with Pass@1 concentrated within ∼.10 of Standard MV plateau (Nemotron3-30B on AIME 2025 and HMMT Feb 2026).

5 Conclusion

We introduced prefix consistency, a reliability signal for weighted majority voting that truncates each CoT and checks whether answers reproduce under regeneration. Across benchmarks, prefix consistency is a stronger correctness predictor than existing baselines, and PC-WMV improves upon existing weighted majority voting methods under cost-equivalent comparison, especially on more difficult benchmarks. Our analysis highlights the discrimination gap $D = r_C - r_W$ as the key quantity governing when the method helps: PC-WMV is most effective when $D > 0$ and Pass@1 leaves a meaningful gap below Standard MV plateau. Regeneration stability is thus a practically useful test-time signal for aggregating votes, not merely a descriptive property of chain-of-thought.

Acknowledgments and Disclosure of Funding

J. Komiyama was supported by the MBZUAI Start-up Fund [BF0121].

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

Pranjal Aggarwal, Seungone Kim, Jack Lanchantin, Sean Welleck, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. OptimalThinkingBench: Evaluating Over and Underthinking in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=N5kWa3sRJt.

Anthropic. Claude Sonnet 4.6 System Card, 2026. URL https://www-cdn.anthropic.com/bbd8ef16d70b7a1665f14f306ee88b53f686aa75.pdf.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on Uncontaminated Math Competitions. In *Thirty-ninth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=y0zL9IZxZ7.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar. 2024. doi: 10.1609/aaai.v38i16.29720. URL https://ojs.aaai.org/index.php/AAAI/article/view/29720.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning Theater: Disentangling Model Beliefs from Chain-of-Thought. *arXiv preprint arXiv:2603.05488*, 2026. URL https://arxiv.org/abs/2603.05488.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT Think That Much for 2+3=?: On the Overthinking of Long Reasoning Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9487–9499. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bx.html.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.558. URL https://aclanthology.org/2024.findings-acl.558/.

Gonçalo Faria and Noah A. Smith. Sample, Don't Search: Rethinking Test-Time Alignment for Language Models. *arXiv preprint arXiv:2504.03790*, 2025. URL https://arxiv.org/abs/2504.03790.

Gonçalo R. A. Faria, Sweta Agrawal, António Farinhas, Ricardo Rei, José G. C. de Souza, and André F. T. Martins. QUEST: Quality-Aware Metropolis-Hastings Sampling for Machine Translation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=dLnduWGTB4.

Yunzhen Feng, Julia Kempe, Cheng Zhang, Parag Jain, and Anthony Hartshorn. What Characterizes Effective Reasoning? Revisiting Length, Review, and Structure of CoT. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025. URL https://openreview.net/forum?id=8VNId3aihj. Spotlight.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-Based Prompting for Multi-step Reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yf1icZHC-l9.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep Think with Confidence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8LqHs0KIM7.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking External Slow-Thinking: From Snowball Errors to Probability of Correct Reasoning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=lAjj22UxZy.

Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. Beyond the Last Answer: Your Reasoning Trace Uncovers More than You Think. *arXiv preprint arXiv:2504.20708*, 2025. URL https://arxiv.org/abs/2504.20708.

Enyi Jiang, Changming Xu, Nischay Singh, Tian Qiu, and Gagandeep Singh. Robust Answers, Fragile Logic: Probing the Decoupling Hypothesis in LLM Reasoning. *arXiv preprint arXiv:2505.17406*, 2025. URL https://arxiv.org/abs/2505.17406.

Ishan Jindal, Sai Prashanth Akuthota, Jayant Taneja, and SACHIN DEV SHARMA. THE PATH OF LEAST RESISTANCE: GUIDING LLM REASONING TRAJECTORIES WITH PREFIX CONSENSUS. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hrnSqERgPn.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Kemp, 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. URL https://arxiv.org/abs/2207.05221.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=29FRqmVQK8.

Juno Kim, Denny Wu, Jason D. Lee, and Taiji Suzuki. Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2HJcVtuovs.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. Best-of-∞: Asymptotic Performance of Test-Time LLM Ensembling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3qiCnLf3jf.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Langton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape Sky-high Cost: Early-stopping Self-Consistency for Multi-step Reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ndR8Ytrzhh.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=HjwK-Tc_Bc.

Mistral AI. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026. URL https://arxiv.org/abs/2601.08584.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

NVIDIA. Nemotron 3 Nano: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning. *arXiv preprint arXiv:2512.20848*, 2025a. URL https://arxiv.org/abs/2512.20848.

NVIDIA. NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model. *arXiv preprint arXiv:2508.14444*, 2025b. URL https://arxiv.org/abs/2508.14444.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Shubham Parashar, Blake Olson, Sambhav Khurana, Eric Li, Hongyi Ling, James Caverlee, and Shuiwang Ji. Inference-Time Computations for LLM Reasoning and Planning: A Benchmark and Insights. *arXiv preprint arXiv:2502.12521*, 2025. URL https://arxiv.org/abs/2502.12521.

Xiao Pu, Michael Saxon, Wenyue Hua, and William Yang Wang. ThoughtTerminator: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=oHR862dpMC.

Daniel Scalena, Leonidas Zotos, Elisabetta Fersini, Malvina Nissim, and Ahmet Üstün. EAGER: Entropy-Aware GEneRation for Adaptive Inference-Time Scaling. *arXiv preprint arXiv:2510.11170*, 2025. URL https://arxiv.org/abs/2510.11170.

Aman Sharma and Paras Chopra. The Sequential Edge: Inverse-Entropy Voting Beats Parallel Self-Consistency at Matched Compute. *arXiv preprint arXiv:2511.02309*, 2025. URL https://arxiv.org/abs/2511.02309.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence Improves Self-Consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are Reasoning LLMs Robust to Interventions on Their Chain-of-Thought? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aQZIpELFwp.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. Knowledge-Driven CoT: Exploring Faithful Reasoning in LLMs for Knowledge-Intensive Question Answering. *arXiv preprint arXiv:2308.13259*, 2023a. URL https://arxiv.org/abs/2308.13259.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170, 2024. doi: 10.1609/aaai.v38i17.29884. URL https://ojs.aaai.org/index.php/AAAI/article/view/29884.

Miles Wang, Robi Lin, Kat Hu, Joy Jiao, Neil Chowdhury, Ethan Chang, and Tejal Patwardhan. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*, 2025. URL https://arxiv.org/abs/2601.21165.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jGbRWwIidy.

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*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. Benchmarking Reasoning Robustness in Large Language Models. *arXiv preprint arXiv:2503.04550*, 2025. URL https://arxiv.org/abs/2503.04550.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O6I0Av7683.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

Jiace Zhu, Yuanzhe Huang, Yingtao Shen, Jie Zhao, and An Zou. Path-Consistency with Prefix Enhancement for Efficient Inference in LLMs. *arXiv preprint arXiv:2409.01281*, 2024. URL `https://arxiv.org/abs/2409.01281`.

Appendix Contents

A Related Work **15**
A.1 Test-Time Scaling . . . 15
A.2 Chain-of-Thought . . . 16

B Limitations and Future Work **16**

C Asymptotic Analysis **17**
C.1 Notation and Useful Tools . . . 17
C.2 Asymptotic Convergence of PC-WMV . . . 19
C.3 Proof of Theorem 1 . . . 20
C.4 Empirical Verification of Assumptions . . . 21

D Additional Experiments **22**
D.1 ROC Curves for Correctness Predictors . . . 22
D.2 Cost–Accuracy Curves . . . 23
D.3 Full Baseline Comparison . . . 23
D.4 Pool Coverage vs. Reweighting . . . 28
D.5 Sensitivity to τ and K . . . 30
D.6 Robustness to Judge Choice . . . 32
D.7 Reliability Signals vs. Pass@1 . . . 34

E Baseline Implementation Details **38**

F Examples of Prefix Consistency **41**

G Evaluation Protocol **42**
G.1 Setup . . . 42
G.2 $\overline{\mathrm{AUROC}}$ for the Correctness-Predictor Evaluation (Section 4.1) . . . 43
G.3 Cost-Accuracy Evaluation (Sections 4.2 and 4.3) . . . 43
G.4 Reproduction-Rate GLM (Section 4.4) . . . 44

H Model Output Postprocessing **45**
H.1 Answer and Verbal Confidence Extraction . . . 45
H.2 LLM-Based Answer Equivalence Judging . . . 46

I Reproducibility Statement **47**

A Related Work

A.1 Test-Time Scaling

Sequential and parallel scaling. Test-time compute scaling comes in two paradigms. Sequential scaling extends a single reasoning chain, for example by appending continuation tokens to lengthen CoT [Muennighoff et al., 2025]. Parallel scaling generates multiple independent solutions and aggregates them: self-consistency is the canonical instance, while tree search [Yao et al., 2023] branches over candidate steps and process-reward verifiers rerank parallel samples [Lightman et al., 2024]. Parashar et al. [2025] benchmarked a range of inference-time techniques across reasoning and planning tasks and reported that no single method is consistently best. Our method sits between the two: it uses parallel samples but introduces a structured perturbation (truncation + regeneration) that probes the robustness of each sample's reasoning. Beyond aggregation, the same perturbation could serve as a quality estimate for partial reasoning states inside richer search procedures such as Graph-of-Thoughts [Besta et al., 2024], which extends Tree-of-Thoughts to a directed graph with refinement and aggregation operations.

Majority voting. Wang et al. [2023b] introduced self-consistency for CoT reasoning: sample multiple reasoning traces and take the majority vote over extracted answers. A line of work targets the cost of self-consistency by stopping early. Aggarwal et al. [2023] introduced Adaptive Consistency, which evaluates a Beta-binomial posterior over the running top-1 and top-2 answer counts and stops once the posterior margin of the top answer crosses a threshold. Li et al. [2024] introduced Early-Stopping Self-Consistency, which inspects fixed-size windows of samples for unanimity and stops at the first unanimous window. Sharma and Chopra [2025] showed that sequential, entropy-aware voting can outperform parallel self-consistency at matched compute, highlighting the importance of cost-equivalent comparisons. Komiyama et al. [2026] analyzed the asymptotic behavior of majority voting (best-of-∞) and proposed a Bayesian nonparametric stopping rule. Our work is complementary: we assess the quality of each vote through regeneration weighting, which is compatible with the adaptive stopping rules above.

Confidence-based weighting. Kang et al. [2025] proposed *self-certainty*, a trace-level quality score defined as the mean KL divergence of the per-token output distribution from uniform, and used it to rerank best-of-N candidates. Fu et al. [2026] then extended self-certainty. In addition to the global trace mean (equivalent to self-certainty), they introduced a sliding-window group confidence and proposed the use of their bottom 10% (bottom-10% group confidence) and the mean over the final tokens of a trace (tail confidence) as alternative trace-level scores for confidence-weighted majority voting. Taubenfeld et al. [2025] demonstrated that combining self-assessed confidence (most effectively via P(True), the model's own probability that its answer is correct) with self-consistency improves accuracy, though this still relies on the model's introspective calibration. Xiong et al. [2024] studied whether LLMs can express uncertainty and found that verbalized confidence is often poorly calibrated. Scalena et al. [2025] used per-token entropy to allocate compute adaptively during generation, a token-level compute-allocation complement to our sample-level reliability estimation. All of these signals rely on the model's own log-probabilities or verbalized self-assessment. We show that such confidence signals are unreliable on difficult problems and propose an alternative based on prefix self-reproduction that requires neither.

Prefix continuation as a per-sample signal. Several recent methods regenerated from a truncated reasoning prefix or extracted a structural signal from a CoT and used it to select or weight samples. Hammoud et al. [2025] proposed SubthoughtReasoner, which segments a single greedy trace into sequential subthoughts at linguistic cues and takes the mode over per-subthought regenerations as the final answer (single-trace refinement, +13pp on AIME 2024 vs. the same trace's last answer). Faria and Smith [2025] also regenerated from intermediate states, but for test-time alignment via Metropolis-Hastings (building on quality-aware machine-translation sampling [Faria et al., 2024]) rather than answer aggregation. Zhu et al. [2024] iteratively promoted a high-confidence prefix among N initial partial answers via Beta-gated agreement on the extracted-answer distribution and resampled subsequent branches conditioned on it (up to 40.5% latency reduction at matched accuracy). Jindal et al. [2026] clustered the first 256 tokens of N initial samples and expanded only the dominant cluster (up to 60% token reduction at matched accuracy). Feng et al. [2025] scored each of $N{=}64$ traces by its Failed-Step Fraction (the proportion of abandoned reasoning branches) and selected the

lowest-FSF trace, yielding 5% to 13% improvements over random selection on AIME 2025. The latter three collapsed N samples onto a subset (a single best prefix, the dominant cluster, or one selected trace) and therefore inherited the failure mode of Standard MV when the subset misses the correct answer: Jindal et al. [2026] reported a 10pp drop on AIME 2025 (76.7% to 66.7%) where the largest prefix cluster need not contain the correct answer. PC-WMV moves in the converse direction: it keeps all N samples and *reweights* the vote by each sample's prefix self-reproduction, so a minority sample whose prefix continuation reproduces its own answer is upweighted rather than discarded. The two directions are complementary, since prefix-cluster pruning could feed PC-WMV reweighting on the surviving traces. Under cost-equivalent comparison, SubthoughtReasoner does not consistently improve over Standard MV in our experiments, filling the comparison left open by Hammoud et al. [2025].

A.2 Chain-of-Thought

Internal answer determination. Several lines of evidence suggest that LLMs determine their answer internally before the visible CoT concludes. Lanham et al. [2023] found that truncating a CoT mid-reasoning and forcing an early answer often does not change the prediction, especially for larger models. Boppana et al. [2026] confirmed this with attention probes: on easy problems, the internal answer is determined well before the visible reasoning ends. Zhang et al. [2025] showed that linear probes on hidden states can classify the correctness of the future final answer at intermediate CoT stages. These observations are consistent with the $r_C > r_W$ asymmetry we measure.

Error propagation. In CoT reasoning, errors compound across steps: once an error is introduced, subsequent reasoning builds on it [Gan et al., 2025]. Kim et al. [2025] modeled CoT as a metastable Markov process on a reasoning graph, where dense intra-cluster (easy) and sparse inter-cluster (difficult) steps induce timescale separation. Our r_C and r_W are analogous in spirit to within-cluster persistence rates, although we measure them at the answer level under regeneration rather than at the reasoning-step level along a single trace.

Reasoning robustness. Several works have studied the robustness of LLM reasoning. Yu et al. [2025] found that reasoning models can be brittle to minor perturbations in the input. von Recum et al. [2026] systematically evaluated seven intervention types on open-weight reasoning LLMs and found that robustness degrades more when interventions occur early in the CoT. Jiang et al. [2025] showed that correct answers persist even when reasoning logic is perturbed, suggesting a decoupling between answer stability and reasoning faithfulness. Our work uses this asymmetry between correct and wrong reasoning traces as a practical signal for answer aggregation.

Overthinking and length scaling. Reasoning models tend to allocate compute disproportionately to problem difficulty, which limits the improvements from extending a single CoT. Chen et al. [2025] reported that o1-like models generate up to 1953% more tokens than non-thinking models on trivial arithmetic, and reach the correct answer in the first generated solution in over 92% of cases while later solutions still account for roughly 40% of tokens. Pu et al. [2025] introduced DUMB500 to quantify overthinking on easy problems and proposed ThoughtTerminator, a training-free decoding-time termination scheme that reduces overthinking tokens by 76% to 98% with minimal accuracy loss. Aggarwal et al. [2026] formalized the trade-off as a joint over- and underthinking benchmark and found that no current model balances the two: even o3 reaches only 71.1% on their unified score. These results motivate aggregating multiple short samples rather than extending a single long trace, which is the setting our method targets.

B Limitations and Future Work

The effectiveness of prefix consistency depends on the discrimination gap $D = r_C - r_W$. When wrong answers are themselves stable under regeneration, r_W remains high, and the advantage of PC-WMV over Standard MV becomes small (e.g., on very difficult problems). Likewise, when the correct answer is reproduced only weakly from the chosen prefix, r_C remains low, and the signal becomes less informative. In this sense, the method is most effective in regimes where regeneration behavior meaningfully separates correct from wrong traces.

The practical advantage of PC-WMV also depends on the available room above Pass@1. When Pass@1 is already close to Standard MV plateau, the aggregated vote on wrong answers is small, so even a strong reliability signal yields only limited accuracy improvements. This explains the small-gap settings in our experiments where PC-WMV does not outperform Standard MV (or simple verbalized-confidence baselines) by a large margin (Appendix D.3).

Although prefix consistency does not require token log-probabilities, it is not without cost: it incurs additional inference cost through prefix truncation and regeneration. This trade-off may be less favorable in deployments where log-probability access is readily available and inexpensive. Relatedly, the method introduces hyperparameters such as the truncation fraction τ, the number of regenerations K, and the weighting function w, and in this work, we use fixed defaults rather than selecting them adaptively for each model or task.

A further practical limitation is that our method assumes access to an explicit Chain-of-Thought trace that can be truncated and continued from a prefix. This assumption does not hold for many frontier closed models and commercial APIs, including systems where the full internal reasoning trace is hidden or only summarized. In such settings, prefix consistency cannot be applied directly, even if the model can generate correct final answers. Extending the method to settings without visible CoT, for example by using intermediate summaries, structured scratchpads, or other externally exposed reasoning states, is an important direction for future work.

More broadly, our results establish prefix consistency as a useful behavioral signal, but not as a mechanistic explanation of why correct traces are more reproducible than wrong ones. Understanding the origin of the observed asymmetry $r_C > r_W$ remains future work. Finally, our evaluation is limited to reasoning-oriented models and math/science-style benchmarks, so the extent to which the same phenomenon holds for other domains, model families, or API settings remains to be established.

As shown in Section 4.4, r_C rises with problem easiness, r_W depends on the model and category but only weakly on Pass@1, and $D = r_C - r_W$ inherits both. This pattern is consistent with reinforcement learning with verifiable rewards (RLVR) encouraging reasoning that survives verification [Wen et al., 2026]. State-of-the-art LLMs are designed to increase the mass of correct reasoning paths via widening the discrimination gap D. We do not test this connection directly, and whether the same pattern holds for non-RLVR-trained reasoners is left to future work.

C Asymptotic Analysis

We analyze the asymptotic convergence of PC-WMV, prove Theorem 1 along the way, and verify the required assumptions empirically.

C.1 Notation and Useful Tools

Assume that the answer space $\mathcal{A}$ is finite throughout this appendix. Fix $\tau \in (0, 1)$ throughout, and omit it from subscripts: e.g., write $\tilde{a}_i$ for $\tilde{a}_i^{(\tau)}$ and $c_i(a)$ for $c_i^{(\tau)}(a)$. The score takes values $c_i(a) \in \{0, \frac{1}{2}, 1\}$.

We define the transition probability as follows:

$$T(b \to a) \;:=\; \Pr\big[\tilde{a}_i = a \,\big|\, a_i = b\big]. \tag{12}$$

The reproduction rates of Section 3.1 can be written in terms of T:

$$r_C \;=\; T(a^\star \to a^\star), \qquad r_W \;=\; \mathbb{E}\big[T(a_i \to a_i) \,\big|\, a_i \neq a^\star\big]\,. \tag{13}$$

Let π and $\pi^{\to}$ be the marginal distributions of a_i and $\tilde{a}_i$, respectively. Here $\pi(a^\star)$ is the same per-problem Pass@1 introduced in Section 2 and used in Theorem 1. We keep the π notation in this appendix so the population-level expressions in Eq. (14) below align with the proof.

$$\pi(a) \;:=\; \Pr[a_i = a], \qquad \pi^{\to}(a) \;:=\; \Pr[\tilde{a}_i = a] \;=\; \sum_{b \in \mathcal{A}} \pi(b)\, T(b \to a). \tag{14}$$

All quantities above (T, r_C, r_W, π, $\pi^{\to}$) are defined per problem. The results in this section are statements about a fixed problem q, and the i.i.d. assumption across i refers to independent samples within q.

For completeness, we restate the definitions of the MV and PC-WMV votes. Let $w\colon [0,1] \to \mathbb{R}_{\geq 0}$ satisfy $w(0) = 0$. For each $a \in \mathcal{A}$, define the aggregated PC-WMV score over N groups by

$$V_N^{(w)}(a) := \sum_{i=1}^{N} w\big(c_i(a)\big),$$

where each term is the per-group PC-WMV vote given in Eq. (8). We then define the PC-WMV estimator and MV voting methods by

$$\hat{a}_N^{\mathrm{PC}} \;:=\; \arg\max_{a\in\mathcal{A}} V_N^{(w)}(a), \qquad \hat{a}_N^{\mathrm{MV}} \;:=\; \arg\max_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{a_i = a\}. \tag{15}$$

We next identify the population objective associated with the PC-WMV estimator.

Proposition 1 (Population objective)**.** *Assume that the pairs $(a_i, \tilde{a}_i)$ are i.i.d. across i, and that, conditional on a_i, the variable $\tilde{a}_i$ is drawn from $T(a_i \to \cdot)$. Then, for every $a \in \mathcal{A}$,*

$$\frac{1}{N} V_N^{(w)}(a) \;\xrightarrow{\text{a.s.}}\; \Phi_w(a) \;:=\; \sum_{b\in\mathcal{A}} \pi(b)\, \mathbb{E}_{Z\sim\mathrm{Bern}(T(b\to a))}\left[w\left(\frac{\mathbf{1}\{a=b\} + Z}{2} \right) \right]. \tag{16}$$

Moreover, if Φ_w has a unique maximizer, then $\hat{a}_N^{\mathrm{PC}}$ converges to that maximizer almost surely.

Proof. For each fixed $a \in \mathcal{A}$, the random variables $\{w(c_i(a))\}_{i=1}^N$ are i.i.d., so the strong law of large numbers yields

$$\frac{1}{N} V_N^{(w)}(a) \;\xrightarrow{\text{a.s.}}\; \mathbb{E}[w(c_i(a))].$$

Conditional on $a_i = b$, we have $\mathbf{1}\{\tilde{a}_i = a\} \sim \mathrm{Bern}(T(b \to a))$, and

$$2c_i(a) = \mathbf{1}\{a = b\} + \mathbf{1}\{\tilde{a}_i = a\}.$$

Hence,

$$\mathbb{E}[w(c_i(a)) \mid a_i = b] = \mathbb{E}_{Z\sim\mathrm{Bern}(T(b\to a))}\left[w\left(\frac{\mathbf{1}\{a=b\} + Z}{2} \right) \right].$$

Taking expectation with respect to $b \sim \pi$ gives Eq. (16). For the final claim, $\frac{1}{N} V_N^{(w)}(a) \to \Phi_w(a)$ a.s. for each a in the finite set $\mathcal{A}$, so when Φ_w has a unique maximizer $a^\star$, eventually $\hat{a}_N^{\mathrm{PC}} = \arg\max_a V_N^{(w)}(a) = a^\star$ almost surely. □

We next derive an explicit decomposition of Φ_w that exposes its dependence on the marginal $\pi(a)$ and the self-transition $T(a \to a)$.

Proposition 2 (Exact decomposition of Φ_w)**.** *For every $a \in \mathcal{A}$,*

$$\Phi_w(a) \;=\; w(\tfrac{1}{2})\left[\pi(a) + \pi^{\to}(a)\right] \;+\; \lambda_w\, \pi(a)\, T(a \to a), \qquad \lambda_w \;:=\; w(1) - 2\, w(\tfrac{1}{2}). \tag{17}$$

Proof. Fix $a \in \mathcal{A}$. Since $c_i(a)$ is the average of the two indicators

$$\mathbf{1}\{a_i = a\} \qquad \text{and} \qquad \mathbf{1}\{\tilde{a}_i = a\},$$

it takes only the three values 0, $\frac{1}{2}$, and 1.

More precisely, $c_i(a) = 1$ if and only if both $a_i = a$ and $\tilde{a}_i = a$ hold. Thus,

$$\Pr\big(c_i(a) = 1\big) = \pi(a)\, T(a \to a).$$

Next, $c_i(a) = \frac{1}{2}$ if and only if exactly one of the two events $\{a_i = a\}$ and $\{\tilde{a}_i = a\}$ occurs. Therefore,

$$\Pr\big(c_i(a) = \tfrac{1}{2}\big) = \Pr(a_i = a, \tilde{a}_i \neq a) + \Pr(a_i \neq a, \tilde{a}_i = a).$$

The first term is

$$\Pr(a_i = a, \tilde{a}_i \neq a) = \pi(a)\big(1 - T(a \to a)\big).$$

For the second term, recall that $\pi^{\rightarrow}(a) = \Pr(\tilde{a}_i = a)$, so

$$\Pr(a_i \neq a, \tilde{a}_i = a) = \Pr(\tilde{a}_i = a) - \Pr(a_i = a, \tilde{a}_i = a) = \pi^{\rightarrow}(a) - \pi(a)T(a \to a).$$

Combining the two expressions yields

$$\Pr\big(c_i(a) = \tfrac{1}{2}\big) = \pi(a) + \pi^{\rightarrow}(a) - 2\pi(a)T(a \to a).$$

Finally, since $w(0) = 0$ and $c_i(a) \in \{0, \frac{1}{2}, 1\}$,

$$\Phi_w(a) = \mathbb{E}[w(c_i(a))] = w(1)\Pr\big(c_i(a) = 1\big) + w(\tfrac{1}{2})\Pr\big(c_i(a) = \tfrac{1}{2}\big).$$

Substituting the above probabilities, we obtain

$$\Phi_w(a) = w(1)\pi(a)T(a \to a) + w(\tfrac{1}{2})\big[\pi(a) + \pi^{\rightarrow}(a) - 2\pi(a)T(a \to a)\big].$$

Rearranging terms gives

$$\Phi_w(a) = w(\tfrac{1}{2})\big[\pi(a) + \pi^{\rightarrow}(a)\big] + \big(w(1) - 2w(\tfrac{1}{2})\big)\pi(a)T(a \to a),$$

which is exactly Eq. (17). □

C.2 Asymptotic Convergence of PC-WMV

Throughout this subsection we assume the i.i.d. setup of Proposition 1.

By Proposition 2, applied to $a^\star$ and to any wrong a and subtracting,

$$\begin{aligned}\Phi_w(a^\star) - \Phi_w(a) \;=\; & \underbrace{w(\tfrac{1}{2})\left[(\pi(a^\star) + \pi^{\rightarrow}(a^\star)) - (\pi(a) + \pi^{\rightarrow}(a))\right]}_{\text{pooled-mass term}} \\ & + \underbrace{\lambda_w\left[\pi(a^\star)\,T(a^\star \to a^\star) - \pi(a)\,T(a \to a)\right]}_{\text{self-reproduction term}}.\end{aligned} \tag{18}$$

By Proposition 1, identifying $a^\star$ as the population maximizer requires this difference to be positive for every wrong a. The two assumptions below control the two terms separately.

Assumption 1 (Self-reproduction dominance)**.** *For every $a \neq a^\star$,*

$$\pi(a^\star)\,T(a^\star \to a^\star) > \pi(a)\,T(a \to a). \tag{19}$$

Interpretation. $\pi(a^\star)\,T(a^\star \to a^\star)$ is the population mass of groups whose initial answer is correct and is reproduced correctly, and $\pi(a)\,T(a \to a)$ is the corresponding mass for the wrong answer a that is reproduced as itself. Assumption 1 requires that correct self-reproduction dominate every individual wrong self-reproduction. Equivalently, it makes the self-reproduction term in the decomposition strictly positive whenever $\lambda_w > 0$.

Assumption 2 (Pooled-mass dominance)**.** *For every $a \neq a^\star$,*

$$\pi(a^\star) + \pi^{\rightarrow}(a^\star) > \pi(a) + \pi^{\rightarrow}(a). \tag{20}$$

Interpretation. $\pi(a) + \pi^{\rightarrow}(a)$ is the total population mass on candidate a, combining initial occurrences and regeneration arrivals. Assumption 2 requires that $a^\star$ have the largest such total. By the same decomposition, it makes the pooled-mass term strictly positive whenever $w(\frac{1}{2}) > 0$.

Theorem 2 (Asymptotic convergence of PC-WMV)**.** *Let w satisfy $w(0) = 0$, $w(1) > 0$, and $\lambda_w := w(1) - 2w(\frac{1}{2}) \geq 0$. Then:*

(a) If Assumptions 1 and 2 hold, then $\hat{a}_N^{\mathrm{PC}} \to a^\star$ almost surely.

(b) If $w(\frac{1}{2}) = 0$, then Assumption 2 is unnecessary: Assumption 1 alone implies $\hat{a}_N^{\mathrm{PC}} \to a^\star$ almost surely.

Proof. Fix any wrong $a \neq a^\star$. By Assumption 1, the self-reproduction bracket of Eq. (18) is strictly positive, so the self-reproduction term is nonnegative and is strictly positive whenever $\lambda_w > 0$.

For part (a), Assumption 2 makes the pooled-mass bracket of Eq. (18) strictly positive, so the pooled-mass term is nonnegative and is strictly positive whenever $w(\frac{1}{2}) > 0$. Since $w(1) = \lambda_w + 2w(\frac{1}{2}) > 0$, at least one of λ_w and $w(\frac{1}{2})$ is strictly positive, so $\Phi_w(a^\star) > \Phi_w(a)$ for every wrong a. Hence $a^\star$ is the unique maximizer of Φ_w, and Proposition 1 yields $\hat{a}_N^{\mathrm{PC}} \to a^\star$ almost surely.

For part (b), if $w(\frac{1}{2}) = 0$ the pooled-mass term vanishes identically and $\lambda_w = w(1) > 0$, so the strict positivity of the self-reproduction term suffices to give $\Phi_w(a^\star) > \Phi_w(a)$ for every wrong a. Thus $\hat{a}_N^{\mathrm{PC}} \to a^\star$ almost surely without Assumption 2. □

C.3 Proof of Theorem 1

We prove Theorem 1 by specializing the exact margin to the binary case $\mathcal{A} = \{a^\star, a'\}$, where $a^\star$ is the correct answer and a' is the only wrong answer (so $\pi(a') = 1 - \pi(a^\star)$). In this case, the two assumptions of Theorem 2 reduce to the same condition and the exact margin takes a one-bracket form, from which the asymptotic boundary $\pi(a^\star) > r_W/(r_C + r_W)$ follows.

The two assumptions coincide. With the only wrong answer being a' and $\pi(a') = 1 - \pi(a^\star)$, Assumption 1 becomes the binary boundary

$$\pi(a^\star)\, T(a^\star \to a^\star) > (1 - \pi(a^\star))\, T(a' \to a'). \tag{21}$$

For Assumption 2, expand $\pi^{\to}$ via Eq. (14), using $T(a^\star \to a') = 1 - T(a^\star \to a^\star)$ and $T(a' \to a^\star) = 1 - T(a' \to a')$ (each row of T sums to one):

$$\pi^{\to}(a^\star) = \pi(a^\star)\, T(a^\star \to a^\star) + (1 - \pi(a^\star))\big(1 - T(a' \to a')\big),$$
$$\pi^{\to}(a') = \pi(a^\star)\big(1 - T(a^\star \to a^\star)\big) + (1 - \pi(a^\star))\, T(a' \to a').$$

Adding $\pi(a^\star)$ and $\pi(a') = 1 - \pi(a^\star)$ and simplifying,

$$\pi(a^\star) + \pi^{\to}(a^\star) = 1 + \big[\pi(a^\star)\, T(a^\star \to a^\star) - (1 - \pi(a^\star))\, T(a' \to a')\big],$$
$$\pi(a') + \pi^{\to}(a') = 1 - \big[\pi(a^\star)\, T(a^\star \to a^\star) - (1 - \pi(a^\star))\, T(a' \to a')\big].$$

Subtracting,

$$\big[\pi(a^\star) + \pi^{\to}(a^\star)\big] - \big[\pi(a') + \pi^{\to}(a')\big] = 2\big[\pi(a^\star)\, T(a^\star \to a^\star) - (1 - \pi(a^\star))\, T(a' \to a')\big].$$

Hence Assumptions 1 and 2 reduce to the same condition Eq. (21). When $\lambda_w \geq 0$, Theorem 2 (a) gives $\hat{a}_N^{\mathrm{PC}} \to a^\star$ a.s. whenever Eq. (21) holds. The exact margin below establishes the result without this restriction on λ_w.

The exact margin. Applying Eq. (18) with $a = a'$, the self-reproduction term contributes λ_w times the quantity $\pi(a^\star)\, T(a^\star \to a^\star) - (1 - \pi(a^\star))\, T(a' \to a')$. By the identity above, the pooled-mass term contributes $2w(\frac{1}{2})$ times the same bracket. Their sum is $w(1) = \lambda_w + 2w(\frac{1}{2})$ times the bracket:

$$\Phi_w(a^\star) - \Phi_w(a') = w(1)\big[\pi(a^\star)\, T(a^\star \to a^\star) - (1 - \pi(a^\star))\, T(a' \to a')\big]. \tag{22}$$

The choice of w thus enters only through $w(1)$.

Proof. By Eq. (22) and $w(1) > 0$, $\Phi_w(a^\star) > \Phi_w(a')$ if and only if

$$\pi(a^\star)\, T(a^\star \to a^\star) > (1 - \pi(a^\star))\, T(a' \to a').$$

By Proposition 1, a positive margin implies $\hat{a}_N^{\mathrm{PC}} \to a^\star$ almost surely, while a negative margin implies convergence to a'. At equality, a group's signed vote is $+w(1)$ or $-w(1)$ with equal positive probabilities, and zero otherwise. Its cumulative sum is a non-degenerate lazy symmetric random walk, which changes sign infinitely often almost surely, so the estimator does not converge to $a^\star$. Identifying $r_C := T(a^\star \to a^\star)$ and $r_W := T(a' \to a')$ via Eq. (13) rewrites the threshold as

$$\pi(a^\star) > \frac{r_W}{r_C + r_W}.$$

Standard MV's vote count $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{a_i = a\}$ converges to $\pi(a)$ a.s. by the strong law of large numbers. At $\pi(a^\star) = \frac{1}{2}$, its signed vote count is a symmetric random walk and does not converge to either answer. Thus Standard MV converges to $a^\star$ almost surely if and only if $\pi(a^\star) > \frac{1}{2}$. Hence PC-WMV converges where Standard MV does not on the interval $r_W/(r_C + r_W) < \pi(a^\star) \leq \frac{1}{2}$, which has positive length if and only if $r_C > r_W$. □

C.4 Empirical Verification of Assumptions

Table 5 checks Assumption 1 (A1) and Assumption 2 (A2) empirically on problems with at least one correct and one wrong initial sample. This is the same observed-sample inclusion rule used for $\overline{\mathrm{AUROC}}$ (Section 4.1, Appendix G.2), and permits estimation of both reproduction rates. All-correct or all-wrong observed pools are excluded because one empirical reproduction rate is unavailable, not because the population probability is known to equal an endpoint. The table further reports the per-problem indicator $\Delta_{w^{(n)}} := \min_{a \neq a^\star} \left[\Phi_{w^{(n)}}(a^\star) - \Phi_{w^{(n)}}(a)\right]$ for $w^{(n)}(c) = c^n$, $n \in \{1, 2, 3\}$ ($\Delta_{w^{(n)}} > 0$ is equivalent to $a^\star$ being the unique maximizer of $\Phi_{w^{(n)}}$).

$\Pr[\mathrm{A1}]$ is non-negligible across a wide range of benchmarks and models, so A1 is not a rare or pathological event in practice. Conditional on A1, $\Pr[\mathrm{A2} \mid \mathrm{A1}]$ is at least 87.5% on every cell and reaches 100% on 40% of them, so the joint occurrence of A1 and A2 is common. For linear weighting ($n = 1$), $\lambda_w = 0$, so $\Delta_{w^{(1)}} > 0$ is exactly A2, explaining the identical columns. For quadratic and cubic weighting ($n \in \{2, 3\}$), the last two columns show that the positive gap occurs with high probability conditional on A1, even when A2 is not required. These findings indicate that the assumptions reflect patterns that arise frequently in the experimental data.

Table 5: **Empirical verification of Theorem 2's assumptions** on $\mathcal{Q}'$, checking A1 and A2 for every wrong a. All probabilities are macro-averaged: each problem in $\mathcal{Q}'$ contributes one indicator per event. Theorem 2 predicts that $a^\star$ uniquely maximizes Φ_w under $\mathrm{A1} \cap \mathrm{A2}$ for every convex nonnegative w with $w(0) = 0$ and $w(1) > 0$. The last three columns report $\Pr[\Delta_{w^{(n)}} > 0 \mid \mathrm{A1}]$ for $w^{(n)}(c) = c^n$, $n \in \{1, 2, 3\}$. At $n = 1$, $\Delta_{w^{(1)}} > 0$ is exactly A2; only $n = 2, 3$ can relax that condition under A1.

Benchmark	Model	$\lvert\mathcal{Q}'\rvert$	$\Pr[\mathrm{A1}]$	$\Pr[\mathrm{A2} \mid \mathrm{A1}]$	$\Pr[\Delta_{w^{(n)}} > 0 \mid \mathrm{A1}]$		
					n=1	n=2	n=3
FrontierScience-Olympiad	GPT-OSS-120B	73	64.4	93.6	93.6	97.9	97.9
	GPT-OSS-20B	74	70.3	96.2	96.2	98.1	100.0
	Nemotron3-30B	80	61.3	93.9	93.9	98.0	100.0
	Nemotron2-9B	59	35.6	90.5	90.5	90.5	100.0
	Ministral3-14B	47	36.2	94.1	94.1	100.0	100.0
HMMT Feb 2026	GPT-OSS-120B	24	79.2	94.7	94.7	100.0	100.0
	GPT-OSS-20B	24	83.3	100.0	100.0	100.0	100.0
	Nemotron3-30B	20	80.0	100.0	100.0	100.0	100.0
	Nemotron2-9B	23	56.5	100.0	100.0	100.0	100.0
	Ministral3-14B	23	69.6	87.5	87.5	93.8	100.0
AIME 2025	GPT-OSS-120B	23	100.0	91.3	91.3	95.7	100.0
	GPT-OSS-20B	26	92.3	100.0	100.0	100.0	100.0
	Nemotron3-30B	17	100.0	100.0	100.0	100.0	100.0
	Nemotron2-9B	23	73.9	100.0	100.0	100.0	100.0
	Ministral3-14B	23	73.9	94.1	94.1	100.0	100.0
Brumo 2025	GPT-OSS-120B	21	81.0	94.1	94.1	94.1	100.0
	GPT-OSS-20B	23	95.7	95.5	95.5	95.5	100.0
	Nemotron3-30B	20	95.0	100.0	100.0	100.0	100.0
	Nemotron2-9B	20	70.0	100.0	100.0	100.0	100.0
	Ministral3-14B	26	88.5	91.3	91.3	95.7	100.0

D Additional Experiments

D.1 ROC Curves for Correctness Predictors

Tables 1 and 2 report $\overline{\text{AUROC}}$ as a single number per (model, benchmark, signal). Figure 4 visualizes the underlying ROC curves on the same 5 model $\times$ 4 benchmark grid, overlaying prefix consistency and the WMV baselines per panel. Curves are macro-averaged: each problem's ROC is interpolated onto a common false-positive-rate grid and then averaged over the same problem subset $\mathcal{Q}'$ used in $\overline{\text{AUROC}}$ (problems with at least one correct and one wrong initial sample, Appendix G.2).

The visual pattern matches the $\overline{\text{AUROC}}$ numbers: prefix consistency's curve sits at or above the baselines on FrontierScience-Olympiad and HMMT Feb 2026 in 8 of the 10 (model, benchmark) settings across the five models, while on AIME 2025 and Brumo 2025 the gap shrinks as the baselines' curves move closer to PC. The two cells where PC is not visibly above are Nemotron2-9B FrontierScience-Olympiad (where P(True) leads, consistent with PC's smallest discrimination gap D in Table 1) and GPT-OSS-120B HMMT Feb 2026 (where DeepConf tail edges PC by .03 in $\overline{\text{AUROC}}$).

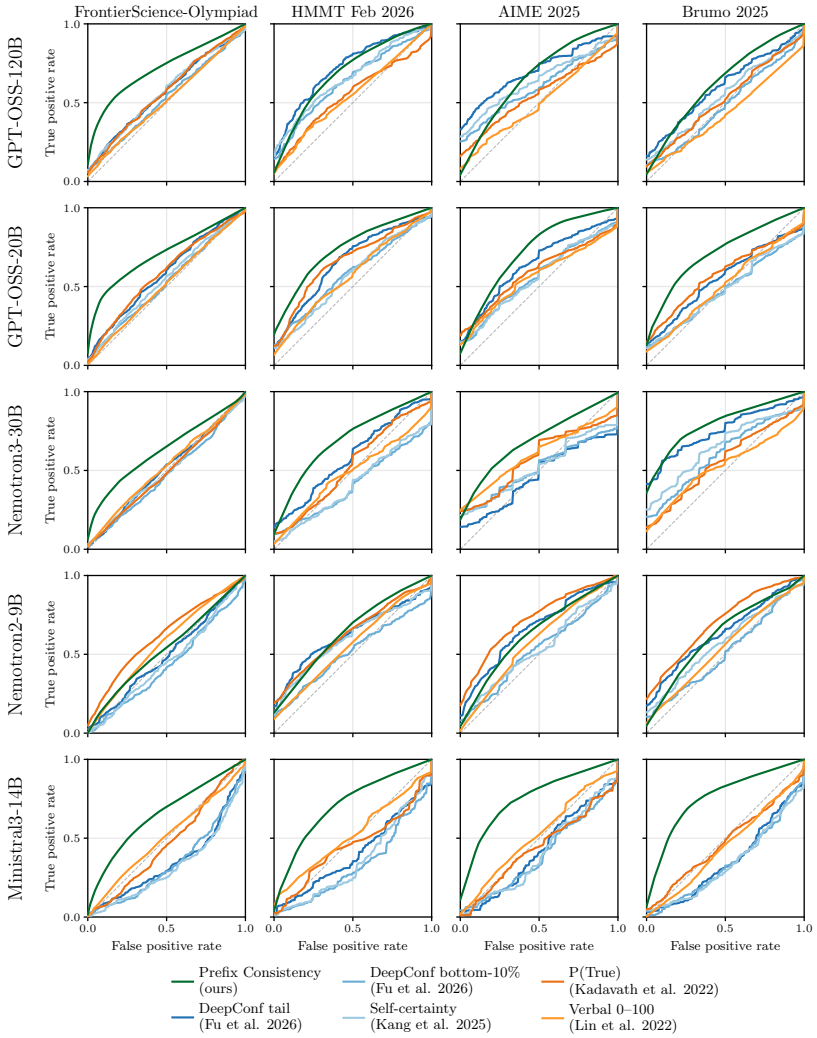


Figure 4: **Macro ROC curves for prefix consistency and the WMV baselines, all 20 (model, benchmark) cells.** Rows are models, columns are benchmarks. Each curve is the per-problem ROC averaged on a common false-positive-rate grid, restricted to problems with at least one correct and one wrong initial sample. The dashed diagonal is the random-classifier baseline.

D.2 Cost–Accuracy Curves

Figure 5 extends the cost–accuracy analysis of Figure 1 to all 20 (model, benchmark) cells, complementing the fixed-budget tables in Appendix D.3. Across nearly all 20 cells PC-cubic (green) sits above the other baselines once the budget exceeds the per-problem regeneration cost, and reaches Standard MV plateau at a noticeably smaller budget.

On FrontierScience-Olympiad, prefix consistency provides stronger correctness discrimination than every competing signal for four of the five models (Table 2). In all four, PC-cubic approaches the Standard MV plateau with substantially fewer tokens. The exception is Nemotron2-9B, which has the smallest discrimination gap (Table 1) and no token savings.

The benefits also extend to settings where conventional confidence signals rank incorrect answers above correct ones. On Ministral3-14B, Self-certainty, DeepConf tail, and P(True) all have below-chance AUROC across the four benchmarks (Table 2). Nevertheless, PC-cubic improves over Standard MV and these baselines on FrontierScience-Olympiad, AIME 2025, and Brumo 2025, despite spending part of the token budget on regeneration.

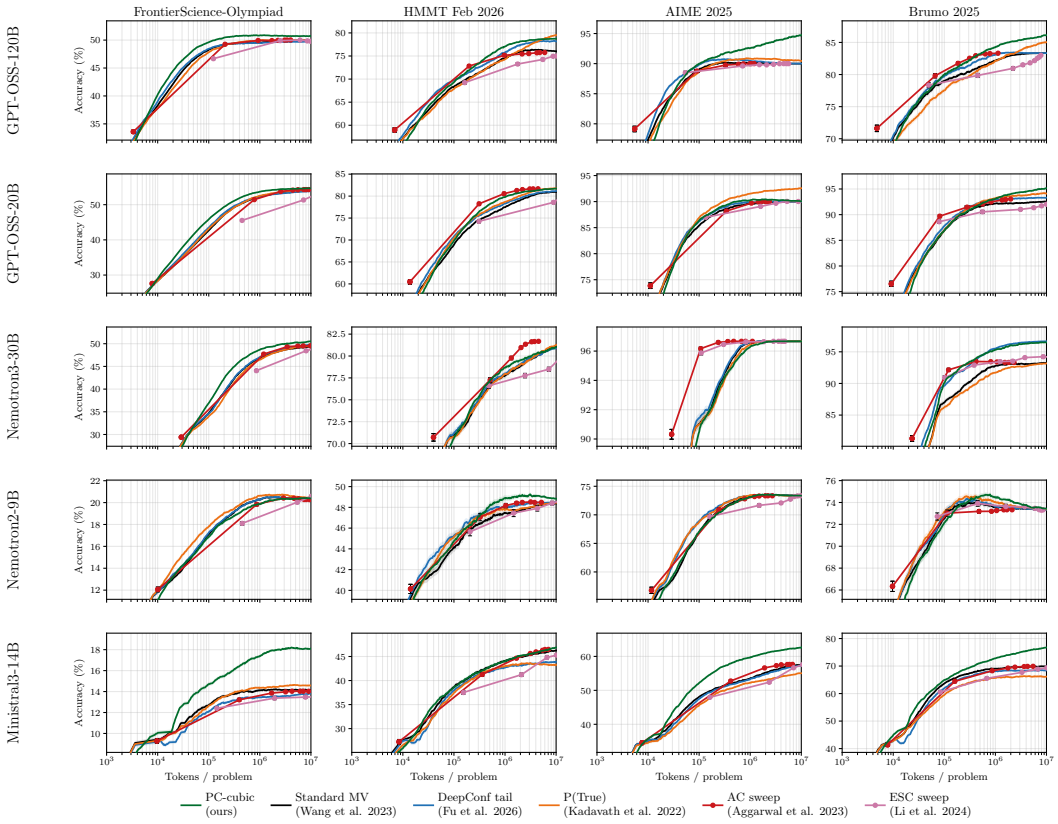


Figure 5: **Cost–accuracy curves for all 20 (model, benchmark) cells.** Rows are models, columns are benchmarks. Y-axes are auto-zoomed to the Pass@1–plateau range. Shaded bands and AC/ESC error bars show $\pm 2\times$ the Monte Carlo standard error of accuracy (vertical) or cost (horizontal).

D.3 Full Baseline Comparison

Figure 6 extends Figure 5 to the full baseline set plus the oracle upper bounds. The three PC variants form a tight green band near the plateau across nearly every cell, while the DeepConf variants and especially their filtered variants spread out widely.

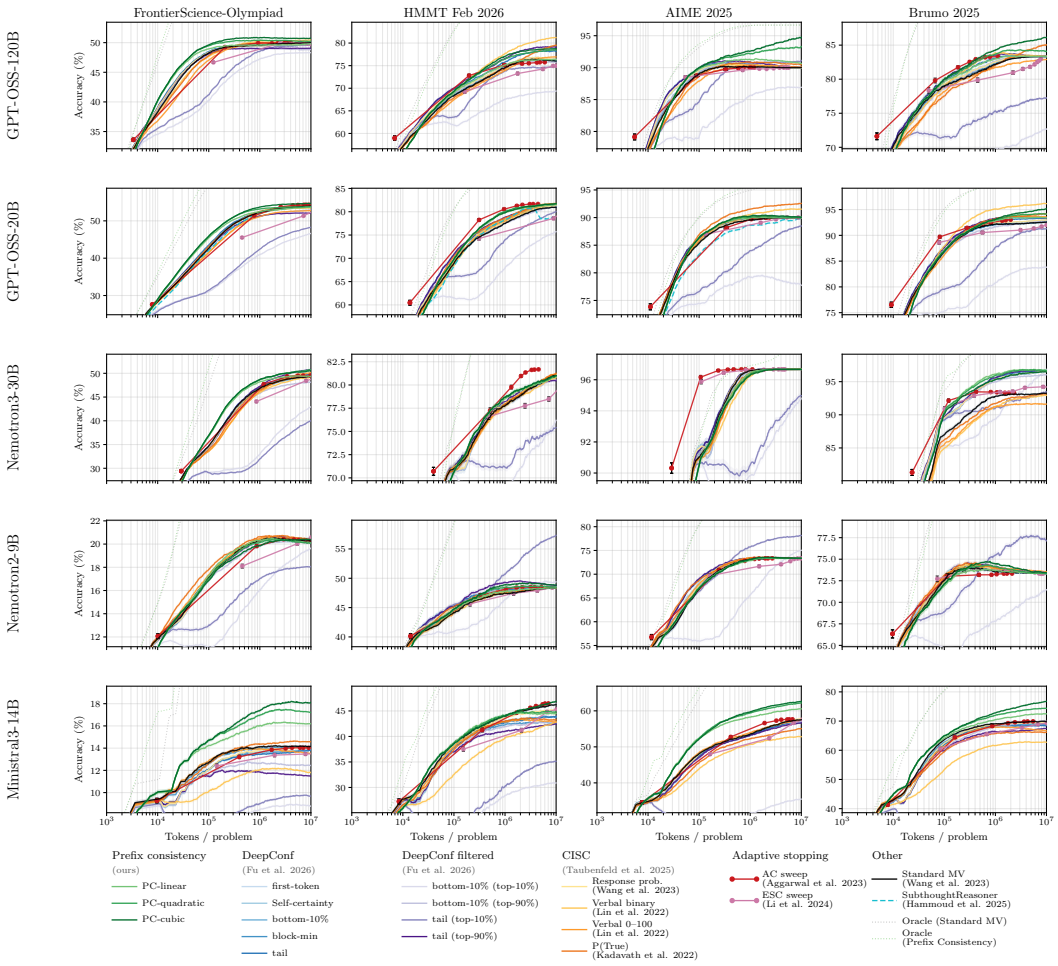


Figure 6: **Cost–accuracy curves with the full baseline set.** Same layout as Figure 5, but additionally includes the oracle upper bounds alongside all evaluated baselines. Methods are colored by family: greens for PC, blues for DeepConf, purples for filtered DeepConf, browns/oranges for CISC, red for AC, mauve for ESC, cyan for SubthoughtReasoner, gray for Standard MV. Oracles are dotted. Shaded bands and AC/ESC error bars show $\pm 2\times$ the Monte Carlo standard error of accuracy (vertical) or cost (horizontal).

Tables 6, 7, 8, 9, and 10 report fixed-budget accuracy at $B \in \{250\text{k}, 1\text{M}, 5\text{M}\}$ tokens for every evaluated baseline (Standard MV, all five DeepConf aggregation strategies [Fu et al., 2026] and their filtered variants, Response probability, verbalized confidence and P(True) [Taubenfeld et al., 2025, Kadavath et al., 2022], and PC-linear/quadratic/cubic), one table per model. See Appendix E for baseline definitions. Tables 11, 12, 13, 14, and 15 report the corresponding token-efficiency ratios at $\alpha \in \{75\%, 90\%, 99\%\}$.

Two patterns emerge. First, on the more difficult science benchmark (FrontierScience-Olympiad) PC-cubic is the best non-oracle method at B=250k and B=1M for the four models with non-trivial discrimination gap (GPT-OSS-120B, GPT-OSS-20B, Nemotron3-30B, Ministral3-14B): at B=1M it reaches .508, .537, .486, and .174 respectively, against the strongest non-PC baseline at .503, .525, .472, and .143. At B=5M PC-cubic remains best on three of these four, with PC-quadratic narrowly ahead on Nemotron3-30B (.504 vs. .503, within 2σ). The PC-linear $\leq$ PC-quadratic $\leq$ PC-cubic ordering holds in 11 of the 12 (model, budget) FrontierScience-Olympiad cells over these four models, with the Nemotron3-30B 5M cell as the only inversion, consistent with the convex-weight analysis:

in Eq. (17), the coefficient $\lambda_w := w(1) - 2w(\frac{1}{2})$ on the self-reproduction term $\pi(a)\,T(a \to a)$ (the joint probability that a is both the initial and the regenerated answer) takes values 0, $\frac{1}{2}$, $\frac{3}{4}$ for PC-linear, PC-quadratic, PC-cubic respectively, so larger n upweights self-reproducing candidates more strongly. The exception is Nemotron2-9B, whose FrontierScience-Olympiad discrimination gap is the smallest in our suite ($D = 7.4\%$, Table 1). PC-WMV there roughly matches Standard MV (PC-cubic .199 vs. Standard MV .203 at B=1M), consistent with Theorem 1 predicting no advantage when D is small.

Second, on the easier benchmarks (AIME 2025, Brumo 2025) where Standard MV plateau already sits close to its peak, verbalized confidence (Verbal binary, P(True)) and DeepConf tail (top-90%) become competitive or take the lead in a few cells (e.g. Verbal binary on GPT-OSS-20B Brumo reaches .951 at B=1M vs. .931 for PC-cubic). This matches the cost–quality trade-off discussed in Section 4, where PC's larger per-group cost is no longer amortized once Standard MV is near its plateau.

SubthoughtReasoner [Hammoud et al., 2025] appears only in the GPT-OSS-20B tables, the only model for which it has been re-evaluated under the unified incremental path. SubthoughtReasoner does not consistently improve over Standard MV (vs. Standard MV at B=1M: .778 vs. .775 on HMMT, .518 vs. .523 on FrontierScience-Olympiad, .886 vs. .896 on AIME 2025), and the GPT-OSS-20B Brumo cell and other-model cells are not evaluated.

Table 6: **All baselines, GPT-OSS-120B (higher accuracy is better).** Bold marks the best non-oracle per column. Subscripts show $\pm 2\times$ the Monte Carlo standard error.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
Baseline												
Standard MV	$.493_{\pm.001}$	$.495_{\pm.001}$	$.499_{\pm.001}$	$.708_{\pm.003}$	$.745_{\pm.003}$	$.763_{\pm.001}$	$.901_{\pm.002}$	$.901_{\pm.001}$	$.900_{\pm.000}$	$.801_{\pm.003}$	$.821_{\pm.002}$	$.833_{\pm.000}$
DeepConf												
DeepConf first-token	$.493_{\pm.001}$	$.495_{\pm.001}$	$.499_{\pm.001}$	$.708_{\pm.003}$	$.745_{\pm.003}$	$.763_{\pm.001}$	$.901_{\pm.002}$	$.901_{\pm.001}$	$.900_{\pm.000}$	$.801_{\pm.003}$	$.821_{\pm.002}$	$.833_{\pm.000}$
Self-certainty	$.494_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.713_{\pm.003}$	$.749_{\pm.003}$	$.768_{\pm.002}$	$.904_{\pm.002}$	$.903_{\pm.001}$	$.900_{\pm.000}$	$.805_{\pm.003}$	$.826_{\pm.002}$	$.833_{\pm.000}$
DeepConf bottom-10%	$.493_{\pm.001}$	$.495_{\pm.001}$	$.496_{\pm.001}$	$.710_{\pm.003}$	$.747_{\pm.003}$	$.764_{\pm.001}$	$.902_{\pm.002}$	$.902_{\pm.001}$	$.900_{\pm.000}$	$.802_{\pm.003}$	$.823_{\pm.002}$	$.833_{\pm.000}$
DeepConf block-min	$.494_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.708_{\pm.003}$	$.747_{\pm.003}$	$.764_{\pm.001}$	$.902_{\pm.002}$	$.902_{\pm.001}$	$.900_{\pm.000}$	$.802_{\pm.003}$	$.824_{\pm.002}$	$.833_{\pm.000}$
DeepConf tail	$.493_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.720_{\pm.003}$	$.758_{\pm.003}$	$.783_{\pm.002}$	$.906_{\pm.002}$	$.905_{\pm.001}$	$.900_{\pm.000}$	$.810_{\pm.003}$	$.831_{\pm.001}$	$.833_{\pm.000}$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$.440_{\pm.002}$	$.472_{\pm.001}$	$.483_{\pm.001}$	$.638_{\pm.004}$	$.675_{\pm.003}$	$.690_{\pm.002}$	$.823_{\pm.003}$	$.856_{\pm.002}$	$.869_{\pm.001}$	$.683_{\pm.004}$	$.697_{\pm.003}$	$.716_{\pm.002}$
DeepConf bottom-10% (top-90%)	$.490_{\pm.001}$	$.490_{\pm.001}$	$.489_{\pm.001}$	$.713_{\pm.003}$	$.746_{\pm.002}$	$.759_{\pm.001}$	$.903_{\pm.002}$	$.903_{\pm.001}$	$.900_{\pm.000}$	$.803_{\pm.003}$	$.825_{\pm.002}$	$.833_{\pm.000}$
DeepConf tail (top-10%)	$.450_{\pm.002}$	$.485_{\pm.001}$	$.495_{\pm.001}$	$.681_{\pm.004}$	$.739_{\pm.003}$	$.745_{\pm.002}$	$.884_{\pm.003}$	$.907_{\pm.002}$	$.909_{\pm.002}$	$.731_{\pm.004}$	$.753_{\pm.004}$	$.770_{\pm.002}$
DeepConf tail (top-90%)	$.489_{\pm.001}$	$.489_{\pm.001}$	$.490_{\pm.001}$	$.728_{\pm.003}$	$.766_{\pm.003}$	$.791_{\pm.002}$	$.909_{\pm.002}$	$.909_{\pm.001}$	$.902_{\pm.001}$	$.816_{\pm.003}$	$.835_{\pm.001}$	$.835_{\pm.001}$
CISC												
Response probability	$.493_{\pm.001}$	$.495_{\pm.001}$	$.497_{\pm.001}$	$.714_{\pm.003}$	$.750_{\pm.003}$	$.768_{\pm.002}$	$.905_{\pm.002}$	$.902_{\pm.001}$	$.900_{\pm.000}$	$.804_{\pm.003}$	$.825_{\pm.002}$	$.833_{\pm.000}$
Verbal binary	$.492_{\pm.001}$	$.503_{\pm.001}$	$.504_{\pm.001}$	$\mathbf{.729}_{\pm.003}$	$\mathbf{.777}_{\pm.002}$	$\mathbf{.807}_{\pm.001}$	$.902_{\pm.002}$	$.904_{\pm.001}$	$.901_{\pm.000}$	$.813_{\pm.003}$	$.839_{\pm.002}$	$.835_{\pm.001}$
Verbal 0–100	$.490_{\pm.001}$	$.499_{\pm.001}$	$.500_{\pm.001}$	$.702_{\pm.003}$	$.739_{\pm.003}$	$.766_{\pm.002}$	$.896_{\pm.002}$	$.904_{\pm.001}$	$.902_{\pm.001}$	$.802_{\pm.003}$	$.818_{\pm.002}$	$.827_{\pm.001}$
P(True)	$.493_{\pm.001}$	$.500_{\pm.001}$	$.500_{\pm.000}$	$.706_{\pm.003}$	$.743_{\pm.003}$	$.784_{\pm.002}$	$.902_{\pm.002}$	$.909_{\pm.001}$	$.906_{\pm.001}$	$.794_{\pm.003}$	$.821_{\pm.003}$	$.844_{\pm.002}$
Prefix consistency												
PC-linear	$.495_{\pm.001}$	$.496_{\pm.001}$	$.496_{\pm.001}$	$.717_{\pm.003}$	$.751_{\pm.002}$	$.764_{\pm.001}$	$.906_{\pm.002}$	$.913_{\pm.002}$	$.909_{\pm.001}$	$.810_{\pm.003}$	$.830_{\pm.002}$	$.833_{\pm.000}$
PC-quadratic	$.502_{\pm.001}$	$.504_{\pm.001}$	$.503_{\pm.001}$	$.725_{\pm.003}$	$.764_{\pm.002}$	$.782_{\pm.001}$	$.912_{\pm.002}$	$.922_{\pm.002}$	$.930_{\pm.002}$	$.815_{\pm.003}$	$.837_{\pm.002}$	$.842_{\pm.001}$
PC-cubic	$\mathbf{.506}_{\pm.001}$	$\mathbf{.508}_{\pm.001}$	$\mathbf{.507}_{\pm.001}$	$.727_{\pm.003}$	$.771_{\pm.002}$	$.786_{\pm.001}$	$\mathbf{.913}_{\pm.002}$	$\mathbf{.926}_{\pm.002}$	$\mathbf{.941}_{\pm.002}$	$\mathbf{.818}_{\pm.003}$	$\mathbf{.842}_{\pm.002}$	$\mathbf{.857}_{\pm.001}$

Table 7: **All baselines, GPT-OSS-20B (higher accuracy is better).** Bold marks the best non-oracle per column. Subscripts show $\pm 2\times$ the Monte Carlo standard error.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
Baseline												
Standard MV	$.482_{\pm.002}$	$.523_{\pm.001}$	$.537_{\pm.001}$	$.736_{\pm.003}$	$.775_{\pm.002}$	$.807_{\pm.001}$	$.881_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.901_{\pm.003}$	$.922_{\pm.002}$	$.924_{\pm.001}$
DeepConf												
DeepConf first-token	$.482_{\pm.002}$	$.523_{\pm.001}$	$.537_{\pm.001}$	$.736_{\pm.003}$	$.775_{\pm.002}$	$.807_{\pm.001}$	$.881_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.901_{\pm.003}$	$.922_{\pm.002}$	$.924_{\pm.001}$
Self-certainty	$.482_{\pm.002}$	$.522_{\pm.001}$	$.535_{\pm.001}$	$.738_{\pm.003}$	$.778_{\pm.003}$	$.808_{\pm.001}$	$.885_{\pm.002}$	$.898_{\pm.001}$	$.900_{\pm.000}$	$.903_{\pm.003}$	$.924_{\pm.002}$	$.927_{\pm.001}$
DeepConf bottom-10%	$.481_{\pm.002}$	$.521_{\pm.001}$	$.536_{\pm.001}$	$.737_{\pm.003}$	$.777_{\pm.002}$	$.808_{\pm.001}$	$.882_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.900_{\pm.003}$	$.921_{\pm.002}$	$.924_{\pm.001}$
DeepConf block-min	$.483_{\pm.002}$	$.523_{\pm.001}$	$.536_{\pm.001}$	$.739_{\pm.003}$	$.776_{\pm.003}$	$.807_{\pm.001}$	$.883_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.900_{\pm.003}$	$.921_{\pm.002}$	$.925_{\pm.001}$
DeepConf tail	$.486_{\pm.002}$	$.524_{\pm.001}$	$.536_{\pm.001}$	$.746_{\pm.003}$	$.783_{\pm.002}$	$.809_{\pm.001}$	$.886_{\pm.002}$	$.899_{\pm.001}$	$.900_{\pm.000}$	$.908_{\pm.003}$	$.930_{\pm.002}$	$.933_{\pm.001}$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$.351_{\pm.002}$	$.410_{\pm.002}$	$.453_{\pm.001}$	$.614_{\pm.004}$	$.678_{\pm.004}$	$.741_{\pm.003}$	$.769_{\pm.004}$	$.792_{\pm.003}$	$.785_{\pm.003}$	$.760_{\pm.004}$	$.808_{\pm.004}$	$.837_{\pm.002}$
DeepConf bottom-10% (top-90%)	$.472_{\pm.002}$	$.512_{\pm.001}$	$.526_{\pm.001}$	$.737_{\pm.003}$	$.777_{\pm.002}$	$.805_{\pm.002}$	$.883_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.896_{\pm.003}$	$.917_{\pm.002}$	$.926_{\pm.001}$
DeepConf tail (top-10%)	$.356_{\pm.002}$	$.425_{\pm.002}$	$.472_{\pm.001}$	$.667_{\pm.005}$	$.740_{\pm.003}$	$.786_{\pm.002}$	$.794_{\pm.004}$	$.840_{\pm.003}$	$.876_{\pm.002}$	$.813_{\pm.005}$	$.876_{\pm.004}$	$.909_{\pm.003}$
DeepConf tail (top-90%)	$.476_{\pm.002}$	$.513_{\pm.002}$	$.521_{\pm.001}$	$.749_{\pm.003}$	$.788_{\pm.002}$	$.811_{\pm.001}$	$.887_{\pm.002}$	$.901_{\pm.002}$	$.900_{\pm.000}$	$.909_{\pm.003}$	$.933_{\pm.002}$	$.936_{\pm.001}$
CISC												
Response probability	$.482_{\pm.002}$	$.522_{\pm.001}$	$.534_{\pm.001}$	$.739_{\pm.003}$	$.778_{\pm.003}$	$.809_{\pm.001}$	$.882_{\pm.002}$	$.897_{\pm.001}$	$.900_{\pm.000}$	$.903_{\pm.003}$	$.924_{\pm.002}$	$.928_{\pm.001}$
Verbal binary	$.465_{\pm.002}$	$.518_{\pm.001}$	$.534_{\pm.001}$	$.742_{\pm.003}$	$.783_{\pm.002}$	$.811_{\pm.001}$	$.885_{\pm.002}$	$.905_{\pm.002}$	$.915_{\pm.002}$	$\mathbf{.922}_{\pm.003}$	$\mathbf{.951}_{\pm.002}$	$\mathbf{.959}_{\pm.001}$
Verbal 0–100	$.465_{\pm.002}$	$.511_{\pm.001}$	$.527_{\pm.001}$	$.741_{\pm.003}$	$.780_{\pm.002}$	$.813_{\pm.001}$	$.887_{\pm.002}$	$.899_{\pm.002}$	$.900_{\pm.001}$	$.903_{\pm.003}$	$.930_{\pm.002}$	$.936_{\pm.001}$
P(True)	$.480_{\pm.002}$	$.525_{\pm.001}$	$.540_{\pm.001}$	$.750_{\pm.003}$	$.787_{\pm.002}$	$\mathbf{.815}_{\pm.001}$	$\mathbf{.895}_{\pm.002}$	$\mathbf{.916}_{\pm.002}$	$\mathbf{.924}_{\pm.001}$	$.906_{\pm.003}$	$.932_{\pm.002}$	$.940_{\pm.002}$
SubthoughtReasoner												
SubthoughtReasoner	$.464_{\pm.002}$	$.518_{\pm.001}$	$.536_{\pm.001}$	$.732_{\pm.004}$	$.778_{\pm.003}$	$.790_{\pm.003}$	$.860_{\pm.003}$	$.886_{\pm.002}$	$.897_{\pm.002}$	—	—	—
Prefix consistency												
PC-linear	$.495_{\pm.002}$	$.524_{\pm.001}$	$.535_{\pm.001}$	$.751_{\pm.003}$	$.797_{\pm.002}$	$.812_{\pm.001}$	$.887_{\pm.002}$	$.899_{\pm.001}$	$.900_{\pm.000}$	$.903_{\pm.003}$	$.928_{\pm.002}$	$.934_{\pm.001}$
PC-quadratic	$\mathbf{.502}_{\pm.002}$	$.532_{\pm.001}$	$.537_{\pm.001}$	$.752_{\pm.003}$	$\mathbf{.800}_{\pm.002}$	$.814_{\pm.001}$	$.888_{\pm.002}$	$.901_{\pm.002}$	$.901_{\pm.001}$	$.904_{\pm.003}$	$.930_{\pm.002}$	$.940_{\pm.001}$
PC-cubic	$\mathbf{.502}_{\pm.002}$	$\mathbf{.537}_{\pm.001}$	$\mathbf{.545}_{\pm.001}$	$\mathbf{.753}_{\pm.004}$	$.797_{\pm.002}$	$.814_{\pm.001}$	$.888_{\pm.002}$	$.902_{\pm.002}$	$.902_{\pm.001}$	$.903_{\pm.003}$	$.931_{\pm.002}$	$.947_{\pm.002}$

Table 8: **All baselines, Nemotron3-30B (higher accuracy is better).** Bold marks the best non-oracle per column. Subscripts show $\pm 2\times$ the Monte Carlo standard error.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
Baseline												
Standard MV	$.404_{\pm.002}$	$.468_{\pm.002}$	$.490_{\pm.001}$	$.736_{\pm.004}$	$.777_{\pm.003}$	$.802_{\pm.002}$	$.935_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.896_{\pm.003}$	$.928_{\pm.002}$	$.931_{\pm.001}$
DeepConf												
DeepConf first-token	$.403_{\pm.002}$	$.467_{\pm.002}$	$.490_{\pm.001}$	$.736_{\pm.004}$	$.777_{\pm.003}$	$.802_{\pm.002}$	$.934_{\pm.002}$	$\mathbf{.966}_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.895_{\pm.003}$	$.927_{\pm.002}$	$.931_{\pm.001}$
Self-certainty	$.402_{\pm.002}$	$.468_{\pm.002}$	$.491_{\pm.001}$	$.734_{\pm.004}$	$.779_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.935_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.918_{\pm.003}$	$.954_{\pm.002}$	$.963_{\pm.001}$
DeepConf bottom-10%	$.401_{\pm.002}$	$.466_{\pm.002}$	$.489_{\pm.001}$	$.732_{\pm.004}$	$.779_{\pm.003}$	$.802_{\pm.002}$	$.930_{\pm.003}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.918_{\pm.003}$	$.953_{\pm.002}$	$.963_{\pm.001}$
DeepConf block-min	$.400_{\pm.002}$	$.467_{\pm.002}$	$.491_{\pm.001}$	$.735_{\pm.004}$	$.778_{\pm.003}$	$.802_{\pm.002}$	$.933_{\pm.003}$	$\mathbf{.966}_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.919_{\pm.003}$	$.954_{\pm.002}$	$.962_{\pm.001}$
DeepConf tail	$.405_{\pm.002}$	$.470_{\pm.002}$	$.494_{\pm.001}$	$.740_{\pm.004}$	$.779_{\pm.003}$	$.801_{\pm.002}$	$\mathbf{.937}_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.920_{\pm.003}$	$.955_{\pm.002}$	$.965_{\pm.001}$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$.297_{\pm.003}$	$.329_{\pm.003}$	$.406_{\pm.002}$	$.688_{\pm.004}$	$.688_{\pm.004}$	$.738_{\pm.004}$	$.889_{\pm.003}$	$.898_{\pm.003}$	$.935_{\pm.002}$	$.860_{\pm.004}$	$.893_{\pm.004}$	$.947_{\pm.003}$
DeepConf bottom-10% (top-90%)	$.398_{\pm.002}$	$.458_{\pm.002}$	$.479_{\pm.002}$	$.732_{\pm.004}$	$.778_{\pm.003}$	$.799_{\pm.002}$	$.930_{\pm.003}$	$.965_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.918_{\pm.003}$	$.950_{\pm.002}$	$.959_{\pm.001}$
DeepConf tail (top-10%)	$.290_{\pm.003}$	$.319_{\pm.002}$	$.381_{\pm.002}$	$.718_{\pm.004}$	$.712_{\pm.004}$	$.746_{\pm.003}$	$.906_{\pm.003}$	$.905_{\pm.003}$	$.939_{\pm.002}$	$.905_{\pm.003}$	$.910_{\pm.003}$	$.928_{\pm.002}$
DeepConf tail (top-90%)	$.404_{\pm.003}$	$.472_{\pm.002}$	$.503_{\pm.002}$	$.740_{\pm.004}$	$.782_{\pm.003}$	$.801_{\pm.002}$	$\mathbf{.937}_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.920_{\pm.003}$	$.955_{\pm.002}$	$.966_{\pm.000}$
CISC												
Response probability	$.401_{\pm.002}$	$.468_{\pm.002}$	$.491_{\pm.001}$	$.734_{\pm.004}$	$.779_{\pm.003}$	$.803_{\pm.002}$	$.934_{\pm.002}$	$\mathbf{.966}_{\pm.000}$	$\mathbf{.967}_{\pm.000}$	$.918_{\pm.003}$	$.954_{\pm.002}$	$.963_{\pm.001}$
Verbal binary	$.386_{\pm.003}$	$.459_{\pm.002}$	$.487_{\pm.002}$	$.734_{\pm.004}$	$.777_{\pm.003}$	$.800_{\pm.002}$	$.925_{\pm.003}$	$.962_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.875_{\pm.003}$	$.912_{\pm.002}$	$.927_{\pm.001}$
Verbal 0–100	$.387_{\pm.003}$	$.462_{\pm.002}$	$.497_{\pm.002}$	$.732_{\pm.004}$	$.779_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.928_{\pm.003}$	$.963_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.878_{\pm.003}$	$.912_{\pm.002}$	$.916_{\pm.001}$
P(True)	$.394_{\pm.003}$	$.465_{\pm.002}$	$.492_{\pm.002}$	$.736_{\pm.004}$	$.779_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.930_{\pm.003}$	$.964_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.886_{\pm.003}$	$.917_{\pm.002}$	$.929_{\pm.001}$
Prefix consistency												
PC-linear	$.429_{\pm.002}$	$.481_{\pm.002}$	$.494_{\pm.001}$	$\mathbf{.742}_{\pm.004}$	$.785_{\pm.003}$	$.802_{\pm.002}$	$.932_{\pm.003}$	$.964_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$\mathbf{.933}_{\pm.003}$	$\mathbf{.961}_{\pm.002}$	$\mathbf{.968}_{\pm.001}$
PC-quadratic	$\mathbf{.433}_{\pm.002}$	$\mathbf{.486}_{\pm.002}$	$\mathbf{.504}_{\pm.002}$	$\mathbf{.742}_{\pm.004}$	$\mathbf{.787}_{\pm.003}$	$.803_{\pm.002}$	$.932_{\pm.003}$	$.963_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.932_{\pm.003}$	$.959_{\pm.002}$	$\mathbf{.968}_{\pm.001}$
PC-cubic	$\mathbf{.433}_{\pm.002}$	$\mathbf{.486}_{\pm.002}$	$.503_{\pm.001}$	$\mathbf{.742}_{\pm.004}$	$\mathbf{.787}_{\pm.003}$	$\mathbf{.804}_{\pm.002}$	$.932_{\pm.003}$	$.963_{\pm.001}$	$\mathbf{.967}_{\pm.000}$	$.929_{\pm.003}$	$.953_{\pm.002}$	$.964_{\pm.001}$

Table 9: **All baselines, Nemotron2-9B (higher accuracy is better).** Bold marks the best non-oracle per column. Subscripts show $\pm 2\times$ the Monte Carlo standard error.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
Baseline												
Standard MV	$.187_{\pm.001}$	$.203_{\pm.001}$	$.203_{\pm.001}$	$.461_{\pm.004}$	$.475_{\pm.002}$	$.482_{\pm.001}$	$.708_{\pm.003}$	$.731_{\pm.002}$	$.734_{\pm.000}$	$.737_{\pm.003}$	$.736_{\pm.002}$	$.734_{\pm.001}$
DeepConf												
DeepConf first-token	$.187_{\pm.001}$	$.203_{\pm.001}$	$.203_{\pm.001}$	$.460_{\pm.004}$	$.475_{\pm.002}$	$.482_{\pm.001}$	$.706_{\pm.003}$	$.730_{\pm.002}$	$.734_{\pm.000}$	$.738_{\pm.003}$	$.736_{\pm.002}$	$.734_{\pm.001}$
Self-certainty	$.187_{\pm.001}$	$.202_{\pm.001}$	$.203_{\pm.001}$	$.468_{\pm.004}$	$.478_{\pm.002}$	$.483_{\pm.001}$	$.717_{\pm.003}$	$.733_{\pm.002}$	$.734_{\pm.001}$	$.743_{\pm.003}$	$.737_{\pm.002}$	$.734_{\pm.001}$
DeepConf bottom-10%	$.186_{\pm.001}$	$.202_{\pm.001}$	$.202_{\pm.001}$	$.463_{\pm.004}$	$.476_{\pm.002}$	$.483_{\pm.001}$	$.709_{\pm.003}$	$.731_{\pm.002}$	$.734_{\pm.001}$	$.739_{\pm.003}$	$.735_{\pm.002}$	$.734_{\pm.001}$
DeepConf block-min	$.188_{\pm.001}$	$.204_{\pm.001}$	$\mathbf{.205}_{\pm.001}$	$.463_{\pm.004}$	$.475_{\pm.002}$	$.482_{\pm.001}$	$.709_{\pm.003}$	$.730_{\pm.002}$	$.734_{\pm.001}$	$.739_{\pm.003}$	$.736_{\pm.002}$	$.734_{\pm.001}$
DeepConf tail	$.188_{\pm.001}$	$.203_{\pm.001}$	$.204_{\pm.001}$	$.471_{\pm.004}$	$.481_{\pm.002}$	$.484_{\pm.000}$	$.716_{\pm.003}$	$.733_{\pm.002}$	$.734_{\pm.001}$	$.741_{\pm.003}$	$.736_{\pm.002}$	$.734_{\pm.001}$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$.127_{\pm.002}$	$.160_{\pm.001}$	$.188_{\pm.001}$	$.422_{\pm.004}$	$.464_{\pm.003}$	$.492_{\pm.003}$	$.568_{\pm.005}$	$.656_{\pm.004}$	$.733_{\pm.003}$	$.654_{\pm.004}$	$.678_{\pm.003}$	$.705_{\pm.003}$
DeepConf bottom-10% (top-90%)	$.183_{\pm.001}$	$.198_{\pm.001}$	$.202_{\pm.001}$	$.467_{\pm.004}$	$.482_{\pm.003}$	$.486_{\pm.001}$	$.707_{\pm.003}$	$.731_{\pm.002}$	$.735_{\pm.001}$	$.739_{\pm.003}$	$.739_{\pm.002}$	$.734_{\pm.001}$
DeepConf tail (top-10%)	$.146_{\pm.002}$	$.170_{\pm.001}$	$.179_{\pm.001}$	$.466_{\pm.004}$	$\mathbf{.506}_{\pm.004}$	$\mathbf{.561}_{\pm.003}$	$.703_{\pm.004}$	$\mathbf{.757}_{\pm.003}$	$\mathbf{.778}_{\pm.002}$	$.716_{\pm.004}$	$\mathbf{.755}_{\pm.004}$	$\mathbf{.778}_{\pm.003}$
DeepConf tail (top-90%)	$.186_{\pm.001}$	$.203_{\pm.001}$	$\mathbf{.205}_{\pm.001}$	$\mathbf{.476}_{\pm.004}$	$.493_{\pm.003}$	$.491_{\pm.001}$	$\mathbf{.719}_{\pm.003}$	$.734_{\pm.002}$	$.735_{\pm.001}$	$.743_{\pm.003}$	$.741_{\pm.002}$	$.735_{\pm.001}$
CISC												
Response probability	$.187_{\pm.001}$	$.202_{\pm.001}$	$.203_{\pm.001}$	$.464_{\pm.004}$	$.477_{\pm.002}$	$.483_{\pm.001}$	$.715_{\pm.003}$	$.734_{\pm.002}$	$.734_{\pm.001}$	$.742_{\pm.003}$	$.737_{\pm.002}$	$.734_{\pm.001}$
Verbal binary	$.189_{\pm.001}$	$.204_{\pm.001}$	$.203_{\pm.001}$	$.461_{\pm.004}$	$.475_{\pm.002}$	$.482_{\pm.001}$	$.707_{\pm.003}$	$.732_{\pm.002}$	$.733_{\pm.000}$	$.743_{\pm.003}$	$.740_{\pm.002}$	$.735_{\pm.001}$
Verbal 0–100	$.189_{\pm.001}$	$.205_{\pm.001}$	$\mathbf{.205}_{\pm.001}$	$.460_{\pm.004}$	$.474_{\pm.002}$	$.483_{\pm.001}$	$.713_{\pm.003}$	$.735_{\pm.002}$	$.734_{\pm.000}$	$.741_{\pm.003}$	$.739_{\pm.002}$	$.737_{\pm.001}$
P(True)	$\mathbf{.192}_{\pm.001}$	$\mathbf{.206}_{\pm.001}$	$\mathbf{.205}_{\pm.001}$	$.468_{\pm.004}$	$.478_{\pm.002}$	$.482_{\pm.001}$	$.713_{\pm.003}$	$.735_{\pm.002}$	$.734_{\pm.000}$	$\mathbf{.745}_{\pm.003}$	$.741_{\pm.002}$	$.735_{\pm.001}$
Prefix consistency												
PC-linear	$.191_{\pm.001}$	$.203_{\pm.001}$	$.203_{\pm.001}$	$.463_{\pm.004}$	$.478_{\pm.002}$	$.483_{\pm.001}$	$.705_{\pm.003}$	$.729_{\pm.002}$	$.733_{\pm.000}$	$.735_{\pm.003}$	$.737_{\pm.002}$	$.734_{\pm.000}$
PC-quadratic	$.188_{\pm.001}$	$.201_{\pm.001}$	$.202_{\pm.001}$	$.466_{\pm.004}$	$.484_{\pm.003}$	$.486_{\pm.001}$	$.705_{\pm.003}$	$.730_{\pm.002}$	$.734_{\pm.000}$	$.738_{\pm.003}$	$.741_{\pm.002}$	$.734_{\pm.000}$
PC-cubic	$.185_{\pm.001}$	$.199_{\pm.001}$	$.204_{\pm.001}$	$.466_{\pm.004}$	$.489_{\pm.003}$	$.490_{\pm.001}$	$.704_{\pm.003}$	$.731_{\pm.002}$	$.734_{\pm.001}$	$.739_{\pm.003}$	$.744_{\pm.002}$	$.735_{\pm.001}$

Table 10: **All baselines, Ministral3-14B (higher accuracy is better).** Bold marks the best non-oracle per column. Subscripts show $\pm 2\times$ the Monte Carlo standard error.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
Baseline												
Standard MV	$.137_{\pm.001}$	$.141_{\pm.001}$	$.142_{\pm.001}$	$.413_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.458_{\pm.002}$	$.508_{\pm.003}$	$.535_{\pm.003}$	$.569_{\pm.002}$	$.669_{\pm.004}$	$.691_{\pm.003}$	$.699_{\pm.003}$
DeepConf												
DeepConf first-token	$.135_{\pm.001}$	$.140_{\pm.001}$	$.142_{\pm.001}$	$.411_{\pm.004}$	$.438_{\pm.003}$	$.458_{\pm.002}$	$.505_{\pm.003}$	$.532_{\pm.003}$	$.569_{\pm.002}$	$.668_{\pm.004}$	$.690_{\pm.003}$	$.698_{\pm.003}$
Self-certainty	$.130_{\pm.001}$	$.135_{\pm.001}$	$.137_{\pm.001}$	$.409_{\pm.004}$	$.429_{\pm.003}$	$.438_{\pm.002}$	$.505_{\pm.003}$	$.533_{\pm.003}$	$.565_{\pm.002}$	$.659_{\pm.004}$	$.682_{\pm.003}$	$.686_{\pm.002}$
DeepConf bottom-10%	$.131_{\pm.001}$	$.136_{\pm.001}$	$.138_{\pm.001}$	$.409_{\pm.004}$	$.430_{\pm.003}$	$.438_{\pm.002}$	$.506_{\pm.003}$	$.531_{\pm.003}$	$.564_{\pm.002}$	$.660_{\pm.004}$	$.683_{\pm.003}$	$.686_{\pm.002}$
DeepConf block-min	$.135_{\pm.001}$	$.140_{\pm.001}$	$.141_{\pm.001}$	$.404_{\pm.004}$	$.423_{\pm.003}$	$.428_{\pm.002}$	$.509_{\pm.003}$	$.534_{\pm.003}$	$.566_{\pm.002}$	$.662_{\pm.004}$	$.683_{\pm.003}$	$.691_{\pm.002}$
DeepConf tail	$.130_{\pm.001}$	$.134_{\pm.001}$	$.137_{\pm.001}$	$.408_{\pm.004}$	$.427_{\pm.003}$	$.436_{\pm.002}$	$.506_{\pm.003}$	$.533_{\pm.003}$	$.565_{\pm.002}$	$.659_{\pm.004}$	$.681_{\pm.003}$	$.683_{\pm.002}$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$.075_{\pm.002}$	$.084_{\pm.001}$	$.089_{\pm.001}$	$.240_{\pm.004}$	$.284_{\pm.003}$	$.304_{\pm.002}$	$.301_{\pm.004}$	$.316_{\pm.004}$	$.347_{\pm.003}$	$.289_{\pm.004}$	$.285_{\pm.003}$	$.262_{\pm.002}$
DeepConf bottom-10% (top-90%)	$.124_{\pm.001}$	$.126_{\pm.001}$	$.125_{\pm.001}$	$.401_{\pm.003}$	$.420_{\pm.003}$	$.428_{\pm.002}$	$.505_{\pm.003}$	$.537_{\pm.003}$	$.567_{\pm.002}$	$.646_{\pm.004}$	$.679_{\pm.003}$	$.696_{\pm.002}$
DeepConf tail (top-10%)	$.079_{\pm.002}$	$.090_{\pm.001}$	$.097_{\pm.001}$	$.249_{\pm.004}$	$.303_{\pm.004}$	$.345_{\pm.002}$	$.276_{\pm.004}$	$.293_{\pm.004}$	$.304_{\pm.002}$	$.303_{\pm.004}$	$.333_{\pm.003}$	$.353_{\pm.002}$
DeepConf tail (top-90%)	$.119_{\pm.001}$	$.119_{\pm.001}$	$.116_{\pm.001}$	$.396_{\pm.003}$	$.410_{\pm.003}$	$.421_{\pm.002}$	$.505_{\pm.003}$	$.532_{\pm.003}$	$.560_{\pm.002}$	$.639_{\pm.004}$	$.664_{\pm.003}$	$.668_{\pm.002}$
CISC												
Response probability	$.132_{\pm.001}$	$.138_{\pm.001}$	$.140_{\pm.001}$	$.410_{\pm.004}$	$.436_{\pm.003}$	$.450_{\pm.002}$	$.508_{\pm.003}$	$.537_{\pm.003}$	$.571_{\pm.002}$	$.663_{\pm.004}$	$.688_{\pm.003}$	$.694_{\pm.002}$
Verbal binary	$.110_{\pm.002}$	$.121_{\pm.001}$	$.120_{\pm.001}$	$.360_{\pm.004}$	$.398_{\pm.003}$	$.418_{\pm.003}$	$.475_{\pm.003}$	$.507_{\pm.003}$	$.525_{\pm.002}$	$.580_{\pm.004}$	$.616_{\pm.003}$	$.627_{\pm.002}$
Verbal 0–100	$.137_{\pm.001}$	$.139_{\pm.001}$	$.136_{\pm.001}$	$.407_{\pm.003}$	$.428_{\pm.003}$	$.430_{\pm.001}$	$.513_{\pm.003}$	$.540_{\pm.003}$	$.568_{\pm.002}$	$.654_{\pm.004}$	$.679_{\pm.003}$	$.672_{\pm.002}$
P(True)	$.137_{\pm.001}$	$.143_{\pm.001}$	$.146_{\pm.001}$	$.404_{\pm.003}$	$.429_{\pm.003}$	$.433_{\pm.002}$	$.495_{\pm.003}$	$.523_{\pm.003}$	$.542_{\pm.002}$	$.632_{\pm.004}$	$.658_{\pm.003}$	$.661_{\pm.003}$
Prefix consistency												
PC-linear	$.149_{\pm.002}$	$.160_{\pm.001}$	$.162_{\pm.001}$	$\mathbf{.418}_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.444_{\pm.002}$	$.556_{\pm.003}$	$.583_{\pm.003}$	$.600_{\pm.002}$	$.675_{\pm.004}$	$.705_{\pm.003}$	$.723_{\pm.002}$
PC-quadratic	$.156_{\pm.002}$	$.170_{\pm.001}$	$.173_{\pm.001}$	$.415_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$.447_{\pm.002}$	$\mathbf{.565}_{\pm.003}$	$.595_{\pm.003}$	$.616_{\pm.002}$	$.684_{\pm.004}$	$.719_{\pm.003}$	$.740_{\pm.002}$
PC-cubic	$\mathbf{.158}_{\pm.002}$	$\mathbf{.174}_{\pm.001}$	$\mathbf{.181}_{\pm.001}$	$.414_{\pm.004}$	$\mathbf{.440}_{\pm.003}$	$\mathbf{.461}_{\pm.002}$	$.564_{\pm.003}$	$\mathbf{.597}_{\pm.003}$	$\mathbf{.621}_{\pm.002}$	$\mathbf{.686}_{\pm.004}$	$\mathbf{.727}_{\pm.003}$	$\mathbf{.759}_{\pm.003}$

Table 11: **Token efficiency, GPT-OSS-120B (smaller is better).** Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates target unreachable. Bold marks the best per column.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
Pass@1 / Standard MV plateau	.338 / .500			.589 / .760			.789 / .900			.713 / .833		
Standard MV budget ($\bar{B}_{\text{MV}}$)	45k	107k	2.5M	364k	858k	1.8M	56k	99k	189k	274k	934k	3.3M
DeepConf												
DeepConf first-token	$1.00_{\pm 0.03}\times$	$1.00_{\pm 0.08}\times$	$1.00_{\pm 0.34}\times$	$1.00_{\pm 0.11}\times$	$1.00_{\pm 0.08}\times$	$1.00_{\pm 0.10}\times$	$1.00_{\pm 0.05}\times$	$1.00_{\pm 0.09}\times$	$1.00_{\pm 0.19}\times$	$1.00_{\pm 0.30}\times$	$1.00_{\pm 0.12}\times$	$1.00_{\pm 0.18}\times$
Self-certainty	$0.89_{\pm 0.04}\times$	$0.87_{\pm 0.06}\times$	N/A	$0.74_{\pm 0.07}\times$	$0.85_{\pm 0.07}\times$	$0.87_{\pm 0.11}\times$	$0.81_{\pm 0.06}\times$	$0.81_{\pm 0.07}\times$	$0.66_{\pm 0.11}\times$	$0.73_{\pm 0.18}\times$	$0.76_{\pm 0.10}\times$	$0.59_{\pm 0.09}\times$
DeepConf bottom-10%	$0.97_{\pm 0.04}\times$	$0.97_{\pm 0.08}\times$	N/A	$0.86_{\pm 0.10}\times$	$0.93_{\pm 0.08}\times$	$0.93_{\pm 0.09}\times$	$0.97_{\pm 0.06}\times$	$0.90_{\pm 0.08}\times$	$0.79_{\pm 0.13}\times$	$0.95_{\pm 0.25}\times$	$0.90_{\pm 0.10}\times$	$0.88_{\pm 0.21}\times$
DeepConf block-min	$0.89_{\pm 0.04}\times$	$0.90_{\pm 0.07}\times$	N/A	$0.93_{\pm 0.11}\times$	$0.96_{\pm 0.08}\times$	$0.92_{\pm 0.09}\times$	$0.98_{\pm 0.06}\times$	$0.95_{\pm 0.08}\times$	$0.86_{\pm 0.14}\times$	$0.95_{\pm 0.23}\times$	$0.88_{\pm 0.10}\times$	$0.96_{\pm 0.21}\times$
DeepConf tail	$0.91_{\pm 0.05}\times$	$0.91_{\pm 0.07}\times$	N/A	$0.63_{\pm 0.09}\times$	$0.72_{\pm 0.08}\times$	$0.57_{\pm 0.07}\times$	$\mathbf{0.68}_{\pm 0.04}\times$	$0.70_{\pm 0.06}\times$	$0.56_{\pm 0.09}\times$	$0.50_{\pm 0.12}\times$	$0.47_{\pm 0.06}\times$	$0.35_{\pm 0.06}\times$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	$>10\times$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
DeepConf bottom-10% (top-90%)	$1.01_{\pm 0.03}\times$	$1.12_{\pm 0.09}\times$	N/A	$0.81_{\pm 0.09}\times$	$0.98_{\pm 0.09}\times$	$1.19_{\pm 0.18}\times$	$0.98_{\pm 0.06}\times$	$0.95_{\pm 0.09}\times$	$0.80_{\pm 0.15}\times$	$0.84_{\pm 0.22}\times$	$0.79_{\pm 0.09}\times$	$0.64_{\pm 0.11}\times$
DeepConf tail (top-10%)	$7.10_{\pm 0.31}\times$	$8.20_{\pm 0.76}\times$	N/A	$1.29_{\pm 0.13}\times$	$1.41_{\pm 0.18}\times$	N/A	$3.14_{\pm 0.21}\times$	$2.87_{\pm 0.27}\times$	$2.12_{\pm 0.34}\times$	N/A	N/A	N/A
DeepConf tail (top-90%)	$0.98_{\pm 0.04}\times$	$1.14_{\pm 0.10}\times$	N/A	$0.45_{\pm 0.04}\times$	$0.52_{\pm 0.05}\times$	$0.42_{\pm 0.04}\times$	$0.69_{\pm 0.04}\times$	$\mathbf{0.69}_{\pm 0.06}\times$	$\mathbf{0.53}_{\pm 0.08}\times$	$0.45_{\pm 0.11}\times$	$0.36_{\pm 0.04}\times$	$0.21_{\pm 0.04}\times$
CISC												
Response probability	$0.95_{\pm 0.05}\times$	$0.96_{\pm 0.08}\times$	N/A	$0.74_{\pm 0.08}\times$	$0.84_{\pm 0.07}\times$	$0.87_{\pm 0.12}\times$	$0.78_{\pm 0.06}\times$	$0.78_{\pm 0.07}\times$	$0.65_{\pm 0.11}\times$	$0.73_{\pm 0.18}\times$	$0.79_{\pm 0.09}\times$	$0.65_{\pm 0.10}\times$
Verbal binary	$1.20_{\pm 0.05}\times$	$1.39_{\pm 0.13}\times$	$0.18_{\pm 0.08}\times$	$0.52_{\pm 0.06}\times$	$\mathbf{0.42}_{\pm 0.03}\times$	$\mathbf{0.31}_{\pm 0.03}\times$	$1.05_{\pm 0.06}\times$	$1.05_{\pm 0.09}\times$	$0.99_{\pm 0.18}\times$	$0.62_{\pm 0.15}\times$	$0.37_{\pm 0.04}\times$	$0.19_{\pm 0.03}\times$
Verbal 0–100	$1.60_{\pm 0.09}\times$	$1.66_{\pm 0.12}\times$	$0.27_{\pm 0.12}\times$	$1.18_{\pm 0.14}\times$	$1.36_{\pm 0.13}\times$	$1.18_{\pm 0.14}\times$	$1.45_{\pm 0.10}\times$	$1.54_{\pm 0.14}\times$	$2.02_{\pm 0.45}\times$	$1.14_{\pm 0.32}\times$	$1.77_{\pm 0.24}\times$	N/A
P(True)	$1.23_{\pm 0.05}\times$	$1.20_{\pm 0.10}\times$	$0.23_{\pm 0.09}\times$	$1.05_{\pm 0.14}\times$	$1.17_{\pm 0.14}\times$	$0.94_{\pm 0.09}\times$	$1.14_{\pm 0.08}\times$	$1.13_{\pm 0.09}\times$	$1.01_{\pm 0.16}\times$	$1.46_{\pm 0.36}\times$	$1.07_{\pm 0.11}\times$	$0.60_{\pm 0.11}\times$
Adaptive stopping												
AC sweep	$1.93_{\pm 0.09}\times$	$1.54_{\pm 0.12}\times$	$0.29_{\pm 0.15}\times$	$\mathbf{0.42}_{\pm 0.05}\times$	$0.63_{\pm 0.10}\times$	N/A	$1.03_{\pm 0.12}\times$	$1.06_{\pm 0.26}\times$	$2.98_{\pm 0.82}\times$	$\mathbf{0.32}_{\pm 0.09}\times$	$\mathbf{0.26}_{\pm 0.03}\times$	$0.20_{\pm 0.04}\times$
ESC sweep	$2.74_{\pm 0.08}\times$	$5.59_{\pm 0.70}\times$	$0.94_{\pm 0.53}\times$	$1.96_{\pm 0.33}\times$	$6.66_{\pm 1.29}\times$	N/A	$0.96_{\pm 0.06}\times$	$1.38_{\pm 0.57}\times$	$>10\times$	$3.19_{\pm 1.47}\times$	$6.11_{\pm 0.65}\times$	N/A
Prefix consistency												
PC-linear	$0.70_{\pm 0.03}\times$	$0.63_{\pm 0.05}\times$	$3.88_{\pm 2.37}\times$	$0.70_{\pm 0.08}\times$	$0.79_{\pm 0.08}\times$	$0.88_{\pm 0.11}\times$	$0.87_{\pm 0.05}\times$	$0.73_{\pm 0.06}\times$	$0.58_{\pm 0.10}\times$	$0.56_{\pm 0.15}\times$	$0.51_{\pm 0.07}\times$	$0.39_{\pm 0.06}\times$
PC-quadratic	$\mathbf{0.65}_{\pm 0.03}\times$	$\mathbf{0.52}_{\pm 0.04}\times$	$0.06_{\pm 0.02}\times$	$0.52_{\pm 0.05}\times$	$0.49_{\pm 0.04}\times$	$0.41_{\pm 0.04}\times$	$0.86_{\pm 0.05}\times$	$0.70_{\pm 0.06}\times$	$\mathbf{0.53}_{\pm 0.08}\times$	$0.44_{\pm 0.11}\times$	$0.37_{\pm 0.04}\times$	$0.19_{\pm 0.03}\times$
PC-cubic	$\mathbf{0.65}_{\pm 0.03}\times$	$\mathbf{0.52}_{\pm 0.04}\times$	$\mathbf{0.05}_{\pm 0.02}\times$	$0.52_{\pm 0.05}\times$	$0.46_{\pm 0.04}\times$	$0.35_{\pm 0.04}\times$	$0.86_{\pm 0.05}\times$	$0.71_{\pm 0.06}\times$	$\mathbf{0.53}_{\pm 0.08}\times$	$0.44_{\pm 0.11}\times$	$0.32_{\pm 0.04}\times$	$\mathbf{0.16}_{\pm 0.02}\times$

Table 12: **Token efficiency, GPT-OSS-20B (smaller is better).** Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates target unreachable. Bold marks the best per column.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
Pass@1 / Standard MV plateau	.276 / .540			.601 / .809			.742 / .900			.766 / .926		
Standard MV budget ($\bar{B}_{\text{MV}}$)	215k	640k	5.0M	517k	1.8M	5.3M	128k	312k	2.1M	150k	348k	5.3M
DeepConf												
DeepConf first-token	$1.00_{\pm 0.03}\times$	$1.00_{\pm 0.07}\times$	$1.00_{\pm 0.15}\times$	$1.00_{\pm 0.11}\times$	$1.00_{\pm 0.09}\times$	$1.00_{\pm 0.18}\times$	$1.00_{\pm 0.09}\times$	$1.00_{\pm 0.16}\times$	$1.00_{\pm 0.24}\times$	$1.00_{\pm 0.06}\times$	$1.00_{\pm 0.13}\times$	$1.00_{\pm 1.33}\times$
Self-certainty	$0.97_{\pm 0.03}\times$	$0.99_{\pm 0.07}\times$	N/A	$0.85_{\pm 0.10}\times$	$0.93_{\pm 0.09}\times$	$0.87_{\pm 0.14}\times$	$0.93_{\pm 0.07}\times$	$0.78_{\pm 0.11}\times$	$0.72_{\pm 0.18}\times$	$0.95_{\pm 0.06}\times$	$0.87_{\pm 0.10}\times$	$0.18_{\pm 0.43}\times$
DeepConf bottom-10%	$1.00_{\pm 0.04}\times$	$1.07_{\pm 0.07}\times$	$1.42_{\pm 0.29}\times$	$0.88_{\pm 0.10}\times$	$0.96_{\pm 0.08}\times$	$0.89_{\pm 0.15}\times$	$0.92_{\pm 0.07}\times$	$0.91_{\pm 0.12}\times$	$0.86_{\pm 0.24}\times$	$1.03_{\pm 0.06}\times$	$1.07_{\pm 0.15}\times$	$0.91_{\pm 1.27}\times$
DeepConf block-min	$0.97_{\pm 0.03}\times$	$1.00_{\pm 0.07}\times$	$1.42_{\pm 0.33}\times$	$0.91_{\pm 0.10}\times$	$0.95_{\pm 0.08}\times$	$0.93_{\pm 0.16}\times$	$0.96_{\pm 0.07}\times$	$0.94_{\pm 0.15}\times$	$0.99_{\pm 0.26}\times$	$1.06_{\pm 0.08}\times$	$1.08_{\pm 0.17}\times$	$0.85_{\pm 1.00}\times$
DeepConf tail	$0.90_{\pm 0.04}\times$	$0.89_{\pm 0.06}\times$	$1.38_{\pm 0.28}\times$	$0.68_{\pm 0.08}\times$	$0.71_{\pm 0.06}\times$	$0.81_{\pm 0.14}\times$	$0.81_{\pm 0.06}\times$	$0.76_{\pm 0.10}\times$	$0.37_{\pm 0.10}\times$	$0.84_{\pm 0.06}\times$	$0.76_{\pm 0.10}\times$	$0.10_{\pm 0.26}\times$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	N/A	N/A	N/A	$>10\times$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
DeepConf bottom-10% (top-90%)	$1.20_{\pm 0.04}\times$	$1.73_{\pm 0.15}\times$	N/A	$0.84_{\pm 0.09}\times$	$1.14_{\pm 0.13}\times$	$1.17_{\pm 0.20}\times$	$0.93_{\pm 0.07}\times$	$0.93_{\pm 0.14}\times$	$0.90_{\pm 0.27}\times$	$1.10_{\pm 0.09}\times$	$1.49_{\pm 0.24}\times$	$0.83_{\pm 1.50}\times$
DeepConf tail (top-10%)	$>10\times$	N/A	N/A	$3.24_{\pm 0.33}\times$	$3.22_{\pm 0.33}\times$	N/A	$>10\times$	$>10\times$	N/A	$8.70_{\pm 0.62}\times$	$>10\times$	N/A
DeepConf tail (top-90%)	$1.11_{\pm 0.04}\times$	$1.65_{\pm 0.16}\times$	N/A	$0.58_{\pm 0.06}\times$	$0.57_{\pm 0.06}\times$	$0.57_{\pm 0.10}\times$	$0.83_{\pm 0.07}\times$	$0.71_{\pm 0.09}\times$	$0.25_{\pm 0.05}\times$	$0.86_{\pm 0.06}\times$	$0.74_{\pm 0.09}\times$	$0.09_{\pm 0.25}\times$
CISC												
Response probability	$0.99_{\pm 0.03}\times$	$1.05_{\pm 0.08}\times$	$2.01_{\pm 0.30}\times$	$0.82_{\pm 0.09}\times$	$0.90_{\pm 0.08}\times$	$0.84_{\pm 0.14}\times$	$1.00_{\pm 0.09}\times$	$0.92_{\pm 0.13}\times$	$0.84_{\pm 0.26}\times$	$0.96_{\pm 0.06}\times$	$0.86_{\pm 0.10}\times$	$0.19_{\pm 0.43}\times$
Verbal binary	$1.40_{\pm 0.06}\times$	$1.25_{\pm 0.08}\times$	N/A	$0.69_{\pm 0.07}\times$	$0.79_{\pm 0.09}\times$	$0.74_{\pm 0.12}\times$	$0.86_{\pm 0.07}\times$	$0.78_{\pm 0.10}\times$	$0.24_{\pm 0.05}\times$	$0.79_{\pm 0.05}\times$	$\mathbf{0.54}_{\pm 0.06}\times$	$\mathbf{0.05}_{\pm 0.14}\times$
Verbal 0–100	$1.37_{\pm 0.06}\times$	$1.76_{\pm 0.13}\times$	N/A	$0.80_{\pm 0.08}\times$	$0.79_{\pm 0.07}\times$	$0.59_{\pm 0.10}\times$	$0.89_{\pm 0.07}\times$	$0.72_{\pm 0.10}\times$	$0.33_{\pm 0.08}\times$	$1.14_{\pm 0.08}\times$	$0.91_{\pm 0.11}\times$	$0.13_{\pm 0.32}\times$
P(True)	$1.00_{\pm 0.04}\times$	$0.94_{\pm 0.05}\times$	$0.56_{\pm 0.08}\times$	$0.61_{\pm 0.07}\times$	$0.63_{\pm 0.07}\times$	$0.53_{\pm 0.09}\times$	$\mathbf{0.73}_{\pm 0.05}\times$	$\mathbf{0.54}_{\pm 0.07}\times$	$\mathbf{0.13}_{\pm 0.03}\times$	$0.97_{\pm 0.07}\times$	$0.79_{\pm 0.09}\times$	$0.10_{\pm 0.26}\times$
SubthoughtReasoner												
SubthoughtReasoner	$1.34_{\pm 0.05}\times$	$1.28_{\pm 0.08}\times$	$1.37_{\pm 0.24}\times$	$0.91_{\pm 0.10}\times$	$0.85_{\pm 0.08}\times$	$0.61_{\pm 0.10}\times$	$2.04_{\pm 0.18}\times$	$3.00_{\pm 0.41}\times$	N/A	—	—	—
Adaptive stopping												
AC sweep	$1.65_{\pm 0.07}\times$	$1.19_{\pm 0.09}\times$	$0.51_{\pm 0.08}\times$	$\mathbf{0.39}_{\pm 0.04}\times$	$\mathbf{0.24}_{\pm 0.03}\times$	$\mathbf{0.21}_{\pm 0.04}\times$	$1.61_{\pm 0.19}\times$	$1.31_{\pm 0.29}\times$	$0.60_{\pm 0.14}\times$	$\mathbf{0.45}_{\pm 0.04}\times$	$0.57_{\pm 0.10}\times$	$0.15_{\pm 0.38}\times$
ESC sweep	$5.04_{\pm 0.41}\times$	$>10\times$	$3.43_{\pm 0.48}\times$	$1.86_{\pm 0.41}\times$	$5.47_{\pm 0.60}\times$	$3.15_{\pm 0.61}\times$	$1.06_{\pm 0.08}\times$	$2.38_{\pm 0.58}\times$	$1.90_{\pm 0.48}\times$	$0.53_{\pm 0.09}\times$	$7.49_{\pm 4.88}\times$	$2.50_{\pm 6.93}\times$
Prefix consistency												
PC-linear	$0.71_{\pm 0.02}\times$	$0.74_{\pm 0.05}\times$	N/A	$0.56_{\pm 0.05}\times$	$0.37_{\pm 0.03}\times$	$0.42_{\pm 0.07}\times$	$0.81_{\pm 0.06}\times$	$0.74_{\pm 0.10}\times$	$0.34_{\pm 0.08}\times$	$0.98_{\pm 0.07}\times$	$0.91_{\pm 0.11}\times$	$0.13_{\pm 0.32}\times$
PC-quadratic	$\mathbf{0.64}_{\pm 0.03}\times$	$\mathbf{0.53}_{\pm 0.03}\times$	$0.89_{\pm 0.38}\times$	$0.55_{\pm 0.05}\times$	$0.36_{\pm 0.03}\times$	$0.28_{\pm 0.05}\times$	$0.81_{\pm 0.06}\times$	$0.72_{\pm 0.10}\times$	$0.26_{\pm 0.06}\times$	$0.96_{\pm 0.06}\times$	$0.89_{\pm 0.10}\times$	$0.12_{\pm 0.29}\times$
PC-cubic	$0.65_{\pm 0.03}\times$	$0.54_{\pm 0.03}\times$	$\mathbf{0.21}_{\pm 0.03}\times$	$0.55_{\pm 0.06}\times$	$0.38_{\pm 0.03}\times$	$0.36_{\pm 0.06}\times$	$0.81_{\pm 0.06}\times$	$0.72_{\pm 0.10}\times$	$0.22_{\pm 0.05}\times$	$0.98_{\pm 0.06}\times$	$0.92_{\pm 0.12}\times$	$0.12_{\pm 0.31}\times$

Table 13: **Token efficiency, Nemotron3-30B (smaller is better).** Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates target unreachable. Bold marks the best per column.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
Pass@1 / Standard MV plateau	.295 / .493			.708 / .810			.902 / .967			.816 / .932		
Standard MV budget ($\bar{B}_{\text{MV}}$)	486k	1.3M	6.0M	1.6M	4.2M	8.5M	396k	589k	1.2M	342k	651k	2.5M
DeepConf												
DeepConf first-token	$1.03_{\pm 0.06}\times$	$1.00_{\pm 0.10}\times$	$0.92_{\pm 0.22}\times$	$1.02_{\pm 0.20}\times$	$1.00_{\pm 0.14}\times$	$0.96_{\pm 0.12}\times$	$1.03_{\pm 0.06}\times$	$1.06_{\pm 0.08}\times$	$0.94_{\pm 0.15}\times$	$1.03_{\pm 0.13}\times$	$1.00_{\pm 0.16}\times$	$0.71_{\pm 0.39}\times$
Self-certainty	$1.04_{\pm 0.06}\times$	$1.01_{\pm 0.11}\times$	$0.85_{\pm 0.19}\times$	$0.85_{\pm 0.15}\times$	$0.89_{\pm 0.14}\times$	$0.93_{\pm 0.12}\times$	$0.97_{\pm 0.06}\times$	$0.96_{\pm 0.06}\times$	$0.84_{\pm 0.12}\times$	$0.45_{\pm 0.05}\times$	$0.41_{\pm 0.05}\times$	$0.14_{\pm 0.07}\times$
DeepConf bottom-10%	$1.07_{\pm 0.05}\times$	$1.11_{\pm 0.11}\times$	$1.36_{\pm 0.34}\times$	$0.93_{\pm 0.18}\times$	$1.00_{\pm 0.15}\times$	$0.97_{\pm 0.13}\times$	$1.01_{\pm 0.06}\times$	$1.00_{\pm 0.06}\times$	$0.90_{\pm 0.15}\times$	$0.50_{\pm 0.06}\times$	$0.41_{\pm 0.05}\times$	$0.15_{\pm 0.08}\times$
DeepConf block-min	$1.05_{\pm 0.05}\times$	$1.01_{\pm 0.11}\times$	$0.85_{\pm 0.20}\times$	$0.91_{\pm 0.18}\times$	$0.98_{\pm 0.15}\times$	$1.00_{\pm 0.13}\times$	$0.97_{\pm 0.05}\times$	$1.00_{\pm 0.07}\times$	$0.90_{\pm 0.15}\times$	$0.44_{\pm 0.06}\times$	$0.40_{\pm 0.05}\times$	$0.15_{\pm 0.08}\times$
DeepConf tail	$0.97_{\pm 0.06}\times$	$0.88_{\pm 0.08}\times$	$0.56_{\pm 0.14}\times$	$0.82_{\pm 0.15}\times$	$0.97_{\pm 0.19}\times$	N/A	$0.95_{\pm 0.06}\times$	$0.95_{\pm 0.06}\times$	$0.89_{\pm 0.13}\times$	$0.40_{\pm 0.05}\times$	$0.39_{\pm 0.07}\times$	$0.14_{\pm 0.07}\times$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	$3.76_{\pm 0.38}\times$	$3.11_{\pm 0.42}\times$	$1.10_{\pm 0.58}\times$
DeepConf bottom-10% (top-90%)	$1.29_{\pm 0.08}\times$	$2.07_{\pm 0.24}\times$	N/A	$1.00_{\pm 0.21}\times$	$1.28_{\pm 0.23}\times$	N/A	$1.02_{\pm 0.06}\times$	$1.03_{\pm 0.07}\times$	$1.05_{\pm 0.17}\times$	$0.50_{\pm 0.06}\times$	$0.42_{\pm 0.06}\times$	$0.16_{\pm 0.08}\times$
DeepConf tail (top-10%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	$0.68_{\pm 0.08}\times$	$2.56_{\pm 0.33}\times$	$2.83_{\pm 1.44}\times$
DeepConf tail (top-90%)	$0.99_{\pm 0.06}\times$	$0.83_{\pm 0.08}\times$	$0.36_{\pm 0.09}\times$	$0.74_{\pm 0.14}\times$	$0.97_{\pm 0.20}\times$	N/A	$0.95_{\pm 0.06}\times$	$0.94_{\pm 0.06}\times$	$0.84_{\pm 0.13}\times$	$0.40_{\pm 0.04}\times$	$0.40_{\pm 0.07}\times$	$0.14_{\pm 0.07}\times$
CISC												
Response probability	$1.04_{\pm 0.06}\times$	$1.01_{\pm 0.11}\times$	$0.86_{\pm 0.21}\times$	$0.83_{\pm 0.15}\times$	$0.90_{\pm 0.15}\times$	$0.94_{\pm 0.12}\times$	$0.97_{\pm 0.05}\times$	$0.96_{\pm 0.06}\times$	$0.84_{\pm 0.13}\times$	$0.45_{\pm 0.06}\times$	$0.41_{\pm 0.05}\times$	$0.15_{\pm 0.08}\times$
Verbal binary	$1.36_{\pm 0.08}\times$	$1.46_{\pm 0.14}\times$	$1.54_{\pm 0.40}\times$	$1.00_{\pm 0.19}\times$	$1.20_{\pm 0.18}\times$	N/A	$1.42_{\pm 0.10}\times$	$1.48_{\pm 0.10}\times$	$1.66_{\pm 0.28}\times$	$1.88_{\pm 0.21}\times$	$2.66_{\pm 0.37}\times$	N/A
Verbal 0–100	$1.34_{\pm 0.08}\times$	$1.18_{\pm 0.11}\times$	$0.59_{\pm 0.15}\times$	$0.89_{\pm 0.16}\times$	$0.89_{\pm 0.13}\times$	$0.86_{\pm 0.11}\times$	$1.15_{\pm 0.07}\times$	$1.25_{\pm 0.11}\times$	$1.15_{\pm 0.18}\times$	$1.86_{\pm 0.24}\times$	N/A	N/A
P(True)	$1.14_{\pm 0.06}\times$	$1.10_{\pm 0.11}\times$	$0.75_{\pm 0.18}\times$	$0.86_{\pm 0.15}\times$	$0.90_{\pm 0.12}\times$	$0.82_{\pm 0.10}\times$	$1.12_{\pm 0.07}\times$	$1.15_{\pm 0.07}\times$	$1.20_{\pm 0.20}\times$	$1.40_{\pm 0.14}\times$	$2.05_{\pm 0.31}\times$	$3.39_{\pm 1.62}\times$
Adaptive stopping												
AC sweep	$1.24_{\pm 0.07}\times$	$0.87_{\pm 0.09}\times$	$0.49_{\pm 0.13}\times$	$\mathbf{0.51}_{\pm 0.09}\times$	$\mathbf{0.34}_{\pm 0.05}\times$	$\mathbf{0.24}_{\pm 0.03}\times$	$\mathbf{0.21}_{\pm 0.01}\times$	$\mathbf{0.18}_{\pm 0.01}\times$	$\mathbf{0.21}_{\pm 0.05}\times$	$\mathbf{0.27}_{\pm 0.03}\times$	$\mathbf{0.18}_{\pm 0.03}\times$	$0.12_{\pm 0.07}\times$
ESC sweep	$2.06_{\pm 0.26}\times$	$3.62_{\pm 0.41}\times$	$2.23_{\pm 0.58}\times$	$4.33_{\pm 1.24}\times$	$3.12_{\pm 0.46}\times$	$1.88_{\pm 0.26}\times$	$0.28_{\pm 0.01}\times$	$0.24_{\pm 0.05}\times$	$0.59_{\pm 0.18}\times$	$0.30_{\pm 0.03}\times$	$0.34_{\pm 0.06}\times$	$0.25_{\pm 0.21}\times$
Prefix consistency												
PC-linear	$0.67_{\pm 0.04}\times$	$0.54_{\pm 0.05}\times$	$0.40_{\pm 0.09}\times$	$0.62_{\pm 0.11}\times$	$0.86_{\pm 0.16}\times$	N/A	$1.13_{\pm 0.08}\times$	$1.21_{\pm 0.08}\times$	$1.29_{\pm 0.24}\times$	$0.34_{\pm 0.04}\times$	$0.28_{\pm 0.04}\times$	$\mathbf{0.09}_{\pm 0.05}\times$
PC-quadratic	$\mathbf{0.61}_{\pm 0.03}\times$	$\mathbf{0.48}_{\pm 0.04}\times$	$0.24_{\pm 0.06}\times$	$0.56_{\pm 0.10}\times$	$0.80_{\pm 0.15}\times$	$1.10_{\pm 0.15}\times$	$1.22_{\pm 0.09}\times$	$1.33_{\pm 0.10}\times$	$1.44_{\pm 0.24}\times$	$0.33_{\pm 0.03}\times$	$0.28_{\pm 0.04}\times$	$0.10_{\pm 0.05}\times$
PC-cubic	$\mathbf{0.61}_{\pm 0.03}\times$	$\mathbf{0.48}_{\pm 0.04}\times$	$\mathbf{0.22}_{\pm 0.05}\times$	$0.59_{\pm 0.10}\times$	$0.79_{\pm 0.13}\times$	$0.95_{\pm 0.14}\times$	$1.23_{\pm 0.08}\times$	$1.39_{\pm 0.10}\times$	$1.67_{\pm 0.26}\times$	$0.35_{\pm 0.04}\times$	$0.30_{\pm 0.04}\times$	$0.11_{\pm 0.06}\times$

Table 14: **Token efficiency, Nemotron2-9B (smaller is better).** Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates target unreachable. Bold marks the best per column.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
Pass@1 / Standard MV plateau	.120 / .203			.401 / .484			.569 / .733			.660 / .733		
Standard MV budget (B_{MV})	181k	383k	841k	274k	1.5M	6.1M	163k	374k	1.1M	77k	109k	126k
DeepConf												
DeepConf first-token	$0.99_{\pm 0.10}\times$	$1.01_{\pm 0.08}\times$	$1.02_{\pm 0.17}\times$	$1.01_{\pm 0.20}\times$	$1.18_{\pm 0.74}\times$	$1.03_{\pm 0.25}\times$	$1.01_{\pm 0.08}\times$	$1.11_{\pm 0.13}\times$	$1.20_{\pm 0.31}\times$	$0.99_{\pm 0.08}\times$	$1.00_{\pm 0.13}\times$	$1.00_{\pm 0.09}\times$
Self-certainty	$1.03_{\pm 0.10}\times$	$1.04_{\pm 0.09}\times$	$1.02_{\pm 0.17}\times$	$0.61_{\pm 0.10}\times$	$0.38_{\pm 0.24}\times$	$0.60_{\pm 0.16}\times$	$0.77_{\pm 0.06}\times$	$0.69_{\pm 0.07}\times$	$0.58_{\pm 0.13}\times$	$0.94_{\pm 0.10}\times$	$0.85_{\pm 0.09}\times$	$\mathbf{0.90}_{\pm 0.08}\times$
DeepConf bottom-10%	$1.07_{\pm 0.11}\times$	$1.12_{\pm 0.10}\times$	$1.14_{\pm 0.18}\times$	$0.82_{\pm 0.14}\times$	$0.66_{\pm 0.37}\times$	$0.91_{\pm 0.22}\times$	$0.95_{\pm 0.07}\times$	$0.90_{\pm 0.10}\times$	$1.00_{\pm 0.26}\times$	$1.02_{\pm 0.08}\times$	$0.95_{\pm 0.11}\times$	$1.04_{\pm 0.11}\times$
DeepConf block-min	$0.95_{\pm 0.09}\times$	$1.01_{\pm 0.11}\times$	$0.94_{\pm 0.15}\times$	$0.86_{\pm 0.17}\times$	$0.87_{\pm 0.43}\times$	$1.31_{\pm 0.33}\times$	$0.95_{\pm 0.07}\times$	$1.00_{\pm 0.13}\times$	$1.33_{\pm 0.32}\times$	$1.05_{\pm 0.08}\times$	$0.98_{\pm 0.12}\times$	$1.05_{\pm 0.10}\times$
DeepConf tail	$0.96_{\pm 0.09}\times$	$1.00_{\pm 0.09}\times$	$1.00_{\pm 0.17}\times$	$0.60_{\pm 0.10}\times$	$0.28_{\pm 0.18}\times$	$0.41_{\pm 0.18}\times$	$0.72_{\pm 0.06}\times$	$0.70_{\pm 0.09}\times$	$0.65_{\pm 0.14}\times$	$0.99_{\pm 0.10}\times$	$0.86_{\pm 0.09}\times$	$0.92_{\pm 0.08}\times$
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	>10×	>10×	N/A	$3.60_{\pm 0.58}\times$	$0.99_{\pm 0.71}\times$	$0.33_{\pm 0.09}\times$	>10×	$8.32_{\pm 0.81}\times$	$4.28_{\pm 0.88}\times$	N/A	N/A	N/A
DeepConf bottom-10% (top-90%)	$1.35_{\pm 0.15}\times$	$1.53_{\pm 0.14}\times$	$2.37_{\pm 0.78}\times$	$0.76_{\pm 0.15}\times$	$0.30_{\pm 0.21}\times$	$0.20_{\pm 0.06}\times$	$0.96_{\pm 0.07}\times$	$0.91_{\pm 0.10}\times$	$0.97_{\pm 0.26}\times$	$1.02_{\pm 0.09}\times$	$0.95_{\pm 0.11}\times$	$1.09_{\pm 0.13}\times$
DeepConf tail (top-10%)	N/A	N/A	N/A	$0.61_{\pm 0.11}\times$	$0.29_{\pm 0.21}\times$	$0.08_{\pm 0.02}\times$	$1.23_{\pm 0.11}\times$	$0.87_{\pm 0.08}\times$	$\mathbf{0.41}_{\pm 0.08}\times$	$3.12_{\pm 0.31}\times$	$3.60_{\pm 0.45}\times$	$3.50_{\pm 0.29}\times$
DeepConf tail (top-90%)	$1.05_{\pm 0.10}\times$	$1.11_{\pm 0.11}\times$	$1.00_{\pm 0.15}\times$	$\mathbf{0.58}_{\pm 0.09}\times$	$\mathbf{0.16}_{\pm 0.11}\times$	$\mathbf{0.07}_{\pm 0.02}\times$	$\mathbf{0.69}_{\pm 0.06}\times$	$\mathbf{0.65}_{\pm 0.07}\times$	$0.57_{\pm 0.13}\times$	$0.99_{\pm 0.10}\times$	$0.86_{\pm 0.09}\times$	$0.94_{\pm 0.09}\times$
CISC												
Response probability	$1.04_{\pm 0.10}\times$	$1.07_{\pm 0.10}\times$	$1.03_{\pm 0.17}\times$	$0.78_{\pm 0.17}\times$	$0.48_{\pm 0.30}\times$	$0.74_{\pm 0.18}\times$	$0.84_{\pm 0.06}\times$	$0.71_{\pm 0.08}\times$	$0.65_{\pm 0.14}\times$	$0.95_{\pm 0.09}\times$	$0.86_{\pm 0.09}\times$	$0.93_{\pm 0.09}\times$
Verbal binary	$0.90_{\pm 0.08}\times$	$1.01_{\pm 0.09}\times$	$0.87_{\pm 0.14}\times$	$1.07_{\pm 0.19}\times$	$0.56_{\pm 0.36}\times$	$1.00_{\pm 0.23}\times$	$1.04_{\pm 0.09}\times$	$1.00_{\pm 0.12}\times$	$0.90_{\pm 0.20}\times$	$0.98_{\pm 0.08}\times$	$1.01_{\pm 0.11}\times$	$1.14_{\pm 0.13}\times$
Verbal 0–100	$0.89_{\pm 0.09}\times$	$0.93_{\pm 0.08}\times$	$0.83_{\pm 0.13}\times$	$1.05_{\pm 0.27}\times$	$0.77_{\pm 0.49}\times$	$0.94_{\pm 0.23}\times$	$0.87_{\pm 0.07}\times$	$0.81_{\pm 0.09}\times$	$0.67_{\pm 0.18}\times$	$0.85_{\pm 0.08}\times$	$0.89_{\pm 0.11}\times$	$0.99_{\pm 0.11}\times$
P(True)	$\mathbf{0.71}_{\pm 0.07}\times$	$\mathbf{0.83}_{\pm 0.09}\times$	$\mathbf{0.65}_{\pm 0.10}\times$	$0.75_{\pm 0.15}\times$	$0.41_{\pm 0.28}\times$	$0.92_{\pm 0.21}\times$	$0.73_{\pm 0.06}\times$	$0.79_{\pm 0.09}\times$	$0.66_{\pm 0.14}\times$	$\mathbf{0.82}_{\pm 0.07}\times$	$0.82_{\pm 0.10}\times$	$\mathbf{0.90}_{\pm 0.09}\times$
Adaptive stopping												
AC sweep	$1.89_{\pm 0.19}\times$	$1.80_{\pm 0.19}\times$	$2.12_{\pm 0.49}\times$	$0.84_{\pm 0.19}\times$	$0.37_{\pm 0.30}\times$	$0.22_{\pm 0.08}\times$	$1.04_{\pm 0.10}\times$	$1.06_{\pm 0.17}\times$	$1.08_{\pm 0.25}\times$	$0.84_{\pm 0.10}\times$	$0.89_{\pm 0.14}\times$	$8.36_{\pm 5.34}\times$
ESC sweep	$2.86_{\pm 0.58}\times$	$6.58_{\pm 1.01}\times$	$7.64_{\pm 1.37}\times$	$1.54_{\pm 0.55}\times$	$1.27_{\pm 1.51}\times$	$1.17_{\pm 0.40}\times$	$1.00_{\pm 0.06}\times$	$4.32_{\pm 1.88}\times$	$6.59_{\pm 1.32}\times$	$0.94_{\pm 0.06}\times$	$\mathbf{0.67}_{\pm 0.13}\times$	$1.31_{\pm 0.46}\times$
Prefix consistency												
PC-linear	$0.84_{\pm 0.08}\times$	$0.86_{\pm 0.08}\times$	$0.83_{\pm 0.12}\times$	$0.90_{\pm 0.17}\times$	$0.41_{\pm 0.27}\times$	$0.94_{\pm 0.36}\times$	$1.08_{\pm 0.10}\times$	$1.03_{\pm 0.12}\times$	$1.23_{\pm 0.26}\times$	$1.19_{\pm 0.13}\times$	$1.37_{\pm 0.18}\times$	$1.47_{\pm 0.16}\times$
PC-quadratic	$0.96_{\pm 0.09}\times$	$1.10_{\pm 0.11}\times$	$1.30_{\pm 0.30}\times$	$0.80_{\pm 0.16}\times$	$0.30_{\pm 0.20}\times$	$0.10_{\pm 0.03}\times$	$1.10_{\pm 0.10}\times$	$1.07_{\pm 0.12}\times$	$1.04_{\pm 0.23}\times$	$1.17_{\pm 0.12}\times$	$1.16_{\pm 0.13}\times$	$1.35_{\pm 0.13}\times$
PC-cubic	$1.11_{\pm 0.11}\times$	$1.46_{\pm 0.17}\times$	$1.80_{\pm 0.31}\times$	$0.78_{\pm 0.14}\times$	$0.28_{\pm 0.19}\times$	$0.09_{\pm 0.02}\times$	$1.12_{\pm 0.09}\times$	$1.08_{\pm 0.12}\times$	$1.00_{\pm 0.20}\times$	$1.17_{\pm 0.12}\times$	$1.16_{\pm 0.14}\times$	$1.34_{\pm 0.13}\times$

Table 15: **Token efficiency, Ministral3-14B (smaller is better).** Cells <1 indicate more cost-efficient than Standard MV. "N/A" indicates target unreachable. Bold marks the best per column.

	FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%	α=75%	α=90%	α=99%
Pass@1 / Standard MV plateau	.091 / .142			.270 / .463			.345 / .576			.421 / .698		
Standard MV budget (B_{MV})	118k	249k	924k	258k	1.2M	7.4M	371k	2.0M	6.9M	92k	265k	1.8M
DeepConf												
DeepConf first-token	$1.10_{\pm 0.10}\times$	$1.24_{\pm 0.37}\times$	$1.27_{\pm 1.04}\times$	$1.09_{\pm 0.13}\times$	$1.24_{\pm 0.28}\times$	$1.00_{\pm 0.30}\times$	$1.24_{\pm 0.15}\times$	$1.04_{\pm 0.11}\times$	$1.02_{\pm 0.20}\times$	$1.04_{\pm 0.08}\times$	$1.01_{\pm 0.16}\times$	$0.96_{\pm 0.48}\times$
Self-certainty	$1.62_{\pm 0.20}\times$	>10×	N/A	$1.29_{\pm 0.17}\times$	N/A	N/A	$1.21_{\pm 0.15}\times$	$1.20_{\pm 0.12}\times$	N/A	$1.35_{\pm 0.11}\times$	$1.47_{\pm 0.22}\times$	N/A
DeepConf bottom-10%	$1.49_{\pm 0.17}\times$	$2.25_{\pm 1.08}\times$	N/A	$1.23_{\pm 0.16}\times$	N/A	N/A	$1.20_{\pm 0.15}\times$	$1.24_{\pm 0.16}\times$	N/A	$1.29_{\pm 0.12}\times$	$1.42_{\pm 0.22}\times$	N/A
DeepConf block-min	$1.09_{\pm 0.11}\times$	$1.39_{\pm 0.39}\times$	$5.93_{\pm 3.20}\times$	$1.81_{\pm 0.31}\times$	N/A	N/A	$0.98_{\pm 0.12}\times$	$1.15_{\pm 0.13}\times$	N/A	$1.16_{\pm 0.09}\times$	$1.42_{\pm 0.24}\times$	N/A
DeepConf tail	$1.54_{\pm 0.18}\times$	>10×	N/A	$1.29_{\pm 0.17}\times$	N/A	N/A	$1.22_{\pm 0.15}\times$	$1.20_{\pm 0.12}\times$	$1.43_{\pm 0.26}\times$	$1.35_{\pm 0.11}\times$	$1.56_{\pm 0.23}\times$	N/A
DeepConf (filtered)												
DeepConf bottom-10% (top-10%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
DeepConf bottom-10% (top-90%)	N/A	N/A	N/A	$2.43_{\pm 0.56}\times$	N/A	N/A	$1.05_{\pm 0.12}\times$	$0.94_{\pm 0.10}\times$	N/A	$1.68_{\pm 0.13}\times$	$2.31_{\pm 0.37}\times$	$2.43_{\pm 1.15}\times$
DeepConf tail (top-10%)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
DeepConf tail (top-90%)	N/A	N/A	N/A	$8.42_{\pm 2.58}\times$	N/A	N/A	$1.28_{\pm 0.19}\times$	$1.63_{\pm 0.25}\times$	N/A	$1.93_{\pm 0.19}\times$	>10×	N/A
CISC												
Response probability	$1.48_{\pm 0.15}\times$	$2.06_{\pm 0.58}\times$	N/A	$1.21_{\pm 0.15}\times$	$1.42_{\pm 0.32}\times$	N/A	$0.97_{\pm 0.11}\times$	$0.93_{\pm 0.10}\times$	$0.86_{\pm 0.15}\times$	$1.24_{\pm 0.10}\times$	$1.26_{\pm 0.20}\times$	$2.45_{\pm 1.64}\times$
Verbal binary	N/A	N/A	N/A	>10×	N/A	N/A	$5.25_{\pm 0.77}\times$	N/A	N/A	>10×	N/A	N/A
Verbal 0–100	$0.91_{\pm 0.10}\times$	$0.97_{\pm 0.27}\times$	N/A	$1.39_{\pm 0.19}\times$	N/A	N/A	$0.90_{\pm 0.13}\times$	$0.96_{\pm 0.11}\times$	$1.29_{\pm 0.22}\times$	$1.49_{\pm 0.12}\times$	$1.61_{\pm 0.24}\times$	N/A
P(True)	$1.10_{\pm 0.12}\times$	$1.01_{\pm 0.27}\times$	$0.48_{\pm 0.30}\times$	$1.33_{\pm 0.14}\times$	N/A	N/A	$1.93_{\pm 0.22}\times$	N/A	N/A	$2.33_{\pm 0.22}\times$	N/A	N/A
Adaptive stopping												
AC sweep	$2.48_{\pm 0.40}\times$	$4.28_{\pm 1.45}\times$	N/A	$1.52_{\pm 0.28}\times$	$1.21_{\pm 0.28}\times$	$\mathbf{0.60}_{\pm 0.20}\times$	$0.91_{\pm 0.11}\times$	$0.54_{\pm 0.06}\times$	$0.49_{\pm 0.10}\times$	$1.42_{\pm 0.14}\times$	$1.85_{\pm 0.32}\times$	$1.20_{\pm 0.75}\times$
ESC sweep	$4.53_{\pm 1.55}\times$	>10×	>10×	$8.79_{\pm 1.33}\times$	$4.75_{\pm 1.03}\times$	$2.05_{\pm 0.75}\times$	$4.44_{\pm 0.83}\times$	$2.41_{\pm 0.22}\times$	$1.38_{\pm 0.26}\times$	$2.39_{\pm 0.38}\times$	$7.42_{\pm 2.14}\times$	$7.95_{\pm 4.31}\times$
Prefix consistency												
PC-linear	$\mathbf{0.38}_{\pm 0.04}\times$	$0.26_{\pm 0.07}\times$	$0.10_{\pm 0.07}\times$	$\mathbf{0.85}_{\pm 0.11}\times$	$1.11_{\pm 0.28}\times$	N/A	$\mathbf{0.25}_{\pm 0.03}\times$	$0.11_{\pm 0.01}\times$	$0.07_{\pm 0.01}\times$	$0.93_{\pm 0.07}\times$	$0.80_{\pm 0.12}\times$	$0.32_{\pm 0.19}\times$
PC-quadratic	$\mathbf{0.38}_{\pm 0.04}\times$	$\mathbf{0.24}_{\pm 0.06}\times$	$\mathbf{0.08}_{\pm 0.05}\times$	$0.91_{\pm 0.11}\times$	$1.01_{\pm 0.26}\times$	N/A	$0.26_{\pm 0.03}\times$	$\mathbf{0.09}_{\pm 0.01}\times$	$\mathbf{0.05}_{\pm 0.01}\times$	$\mathbf{0.86}_{\pm 0.06}\times$	$\mathbf{0.61}_{\pm 0.09}\times$	$0.20_{\pm 0.12}\times$
PC-cubic	$\mathbf{0.38}_{\pm 0.04}\times$	$\mathbf{0.24}_{\pm 0.06}\times$	$\mathbf{0.08}_{\pm 0.05}\times$	$1.03_{\pm 0.12}\times$	$0.98_{\pm 0.20}\times$	$0.67_{\pm 0.20}\times$	$0.27_{\pm 0.03}\times$	$\mathbf{0.09}_{\pm 0.01}\times$	$\mathbf{0.05}_{\pm 0.01}\times$	$\mathbf{0.86}_{\pm 0.06}\times$	$0.62_{\pm 0.09}\times$	$\mathbf{0.18}_{\pm 0.11}\times$

D.4 Pool Coverage vs. Reweighting

We show that PC-WMV's advantage over the non-PC baselines comes from reweighting, not from an enlarged pool of correct candidates.

Oracle upper bounds. We define Oracle (Standard MV) as the accuracy of an ideal selector that, at each budget B, always picks the correct answer from the initial-sample pool whenever one exists, and Oracle (Prefix Consistency) as the same ideal selector applied to the combined pool of initial and regenerated answers. Both oracles are unattainable in practice, since the correct answer is unknown at inference time; they serve as upper bounds on what Standard MV and PC-WMV can achieve, respectively. Figure 7 replots Figure 6 restricted to Standard MV, PC-cubic, and the two oracle upper bounds, with the y-axis chosen so the oracles stay in view at high token budgets. Oracle (Standard MV) and Oracle (Prefix Consistency) coincide within $\pm 2\times$ the Monte Carlo standard error on most of the 20 cells, and the residual gap between them is small relative to the PC-cubic vs. Standard MV gap in the same cells. At a fixed token budget B, the union of correct candidates reachable from sampled groups is therefore essentially the same as that reachable from the equivalent number of initial samples drawn under the Standard MV protocol. Three factors related to pool coverage explain the small residuals. First, at τ=0.75, K=1 each PC group costs $\approx 1.25\times$ a Standard MV sample,[2] so at budget B PC pools roughly $1.6\times$ as many candidate answers as Standard MV. Second, $\tilde{a}_i$ is correlated with a_i through the shared prefix $y_i[:\lceil \tau |y_i| \rceil]$, so the effective number of independent answers in the group pool is well below $1.6\times$, nearly canceling the first effect. Third, the asymptotic pool of correct candidates is by construction weakly larger for the group pool than for the initial pool,

[2]The factor $1.25 = 1 + (1 - \tau)$ is the truncation-implied expectation. Continuation/generation token ratios in Table 26 give empirical per-group cost factors of 1.25–$1.34\times$ on four of the five models and 1.49–$1.60\times$ on Ministral3-14B.

which puts Oracle (Prefix Consistency) weakly above Oracle (Standard MV) at high B in a subset of cells. None of these effects involves the reweighting, so they cannot account for the PC-cubic vs. Standard MV gap in the same panels. PC-WMV's advantage therefore comes from the weighting itself, with the prefix-consistency score $c_i(a)$ acting as a per-candidate reliability signal (Section 3.1).

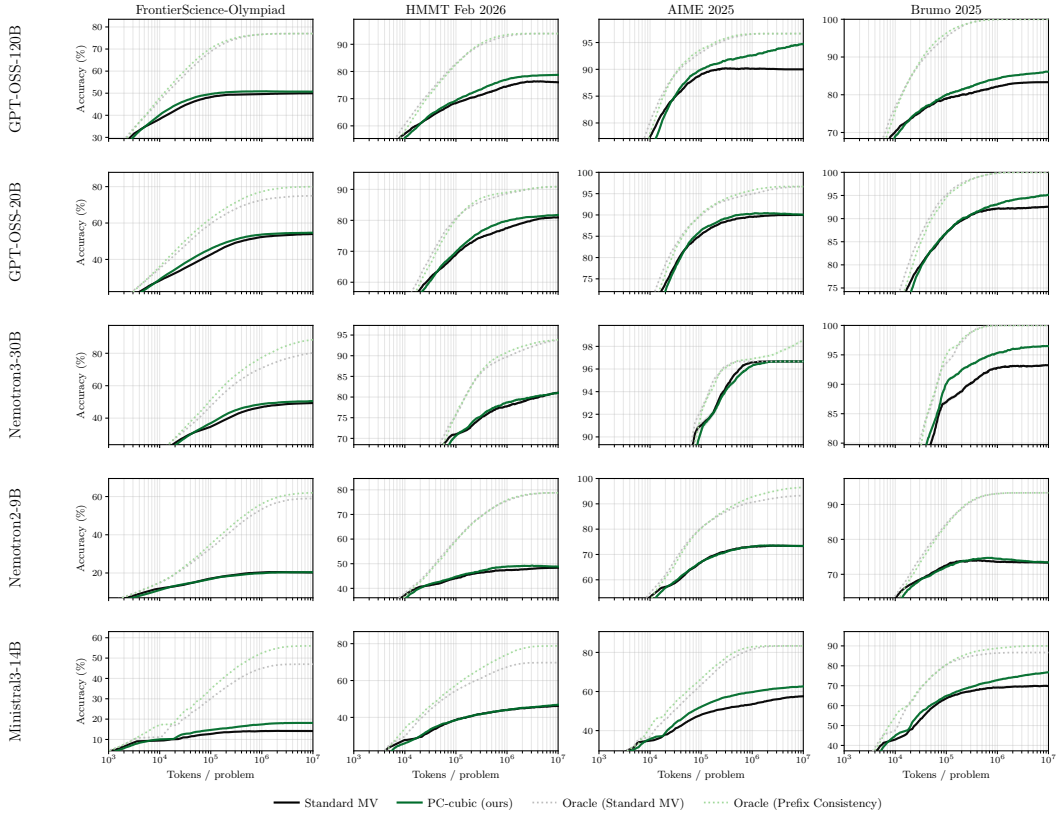


Figure 7: **Oracle decomposition: pool coverage vs. reweighting.** Same 5x4 layout as Figure 6, restricted to Standard MV, PC-cubic, and the two oracle upper bounds. Y-axes include the oracle range so the high-budget tails stay in view. The two oracles nearly coincide in every cell, indicating that PC-WMV's advantage over the non-PC baselines does not come from regeneration adding new correct candidates to the pool. Shaded bands show $\pm 2\times$ the Monte Carlo standard error of accuracy.

Per-problem marginal correctness. Figure 8 gives a complementary per-problem view. Each point is the empirical estimate of the per-problem marginals $\pi(a^\star) = \Pr[a_i = a^\star]$ over the initial answers and $\pi^{\rightarrow}(a^\star) = \Pr[\tilde{a}_i = a^\star]$ over the regenerations (notation as defined in Appendix C.1, Eq. (14)). The plotted estimates $\hat{\pi}(a^\star)$ and $\widehat{\pi^{\rightarrow}}(a^\star)$ cluster tightly on the $y = x$ diagonal in every cell, and their cell-level means agree within a few percentage points on most cells. The largest cell-mean difference is for Nemotron3-30B on AIME 2025, where the cell mean of $\hat{\pi}(a^\star)$ is 0.90 and that of $\widehat{\pi^{\rightarrow}}(a^\star)$ is 0.77, and across the 20 cells the differences run in both directions (e.g., Ministral3-14B has the cell mean of $\widehat{\pi^{\rightarrow}}(a^\star)$ above that of $\hat{\pi}(a^\star)$ on AIME 2025 and Brumo 2025). A paired Wilcoxon signed-rank test on the per-problem differences $\hat{\pi}_q(a^\star) - \widehat{\pi^{\rightarrow}}_q(a^\star)$ rejects equality at the Bonferroni-corrected significance level $\alpha_{\text{sig}} = 0.05/20$ on 5 of the 20 cells, with absolute cell-mean differences of at most 0.126 (the Nemotron3-30B AIME 2025 outlier, with the other four cells within 0.061). Among these 5 detected cells, 4 have the cell mean of $\hat{\pi}(a^\star)$ above the cell mean of $\widehat{\pi^{\rightarrow}}(a^\star)$, that is, regenerated answers are statistically less correct on average than initial answers (the only exception is Ministral3-14B FrontierScience-Olympiad). Across all 20 cells the cell-mean difference is positive on 13 cells and negative on 7, indistinguishable from a $50/50$ split (two-sided

binomial sign test, $p = 0.26$). The marginal correctness rate is therefore comparable for initial and regenerated answers, and the cell-mean differences do not systematically favor the regenerations. This is a stronger statement than mere equality of rates: PC-WMV's advantage cannot come from regeneration raising the per-answer correctness rate, because in the cells where the rate differs the regenerations have the lower marginal correctness.

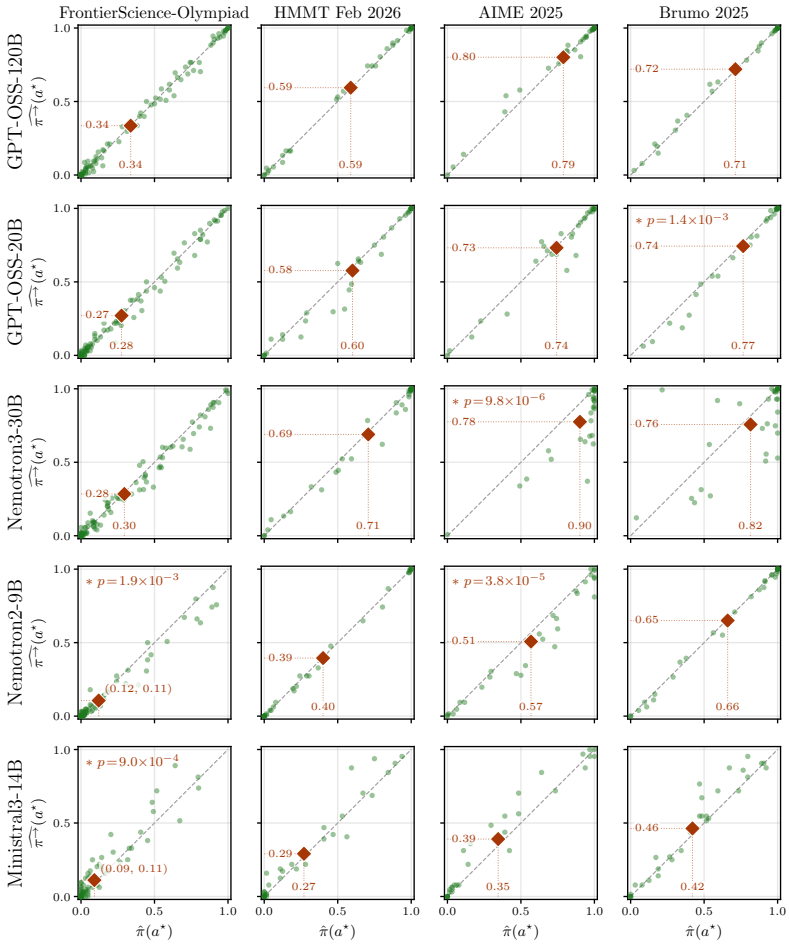


Figure 8: **Per-problem $\hat{\pi}(a^\star)$ vs. $\widehat{\pi^{\rightarrow}}(a^\star)$.** Each point is one problem and the dashed line marks $y = x$. Red diamonds with dotted projection lines show each cell's mean and its coordinates. $*$ flags cells where a paired Wilcoxon test rejects equality (Bonferroni-corrected significance level $\alpha_{\text{sig}} = 0.05/20$, not to be confused with the target-accuracy ratio α of Section 4.3, with the panel-level p-value shown).

D.5 Sensitivity to τ and K

Section 3.1 fixes $K = 1$ to keep notation succinct. Here we extend it to the general-K multiset of Eq. (4) and the corresponding score $c_i^{(\tau,K)}(a) = |\{a' \in A_i^{(\tau,K)} : a' = a\}|/(K+1)$, sweeping $K \in \{1, 2, 3\}$ and $\tau \in \{0.25, 0.50, 0.75\}$.

Table 16 sweeps the truncation fraction $\tau \in \{0.25, 0.50, 0.75\}$ at K=1 for GPT-OSS-20B. Both r_C and r_W decrease with deeper truncation, with r_W generally decreasing faster (e.g., on AIME 2025, r_C goes $87.8\% \to 73.0\%$ and r_W goes $42.6\% \to 17.1\%$ as τ moves $0.75 \to 0.25$). The discrimination gap $D(\tau)$ varies by benchmark: it widens monotonically on AIME 2025 ($45.2 \to 50.0 \to 55.9$), peaks at τ=0.50 on HMMT and Brumo 2025, and narrows on FrontierScience-Olympiad ($41.4 \to 39.5 \to 39.1$). Deeper truncation also costs more per group, since regeneration covers a longer suffix. Cost-equivalent accuracy across (τ, K) is reported in Table 17.

Table 16: **Reproduction rates (%) across truncation fractions $\tau \in \{0.75, 0.50, 0.25\}$ at $K{=}1$, GPT-OSS-20B (larger D is better).** Column symbols (r_C, r_W, D) follow Table 1, with $\overline{\text{AUROC}} = (1 + D)/2$ at $K{=}1$. Macro-averaged over problems with at least one correct and one wrong initial sample.

Model	Benchmark	τ	r_C	r_W	D	$\overline{\text{AUROC}}$
GPT-OSS-20B	FrontierScience-Olympiad	0.75	**55.4**	**14.0**	**41.4**	0.707
		0.50	**48.1**	**8.6**	**39.5**	0.697
		0.25	**45.0**	**5.9**	**39.1**	0.696
GPT-OSS-20B	HMMT Feb 2026	0.75	**78.3**	**29.7**	**48.6**	0.743
		0.50	**67.8**	**18.1**	**49.7**	0.748
		0.25	**60.2**	**12.3**	**47.9**	0.740
GPT-OSS-20B	AIME 2025	0.75	**87.8**	**42.6**	**45.2**	0.726
		0.50	**78.6**	**28.6**	**50.0**	0.750
		0.25	**73.0**	**17.1**	**55.9**	0.780
GPT-OSS-20B	Brumo 2025	0.75	**82.8**	**39.0**	**43.7**	0.719
		0.50	**74.8**	**18.3**	**56.5**	0.782
		0.25	**67.6**	**16.5**	**51.1**	0.756

Table 17: **Sensitivity of PC-WMV accuracy to (τ, K) on GPT-OSS-20B (higher is better).** Accuracy of PC-linear, PC-quadratic, PC-cubic at budgets 250k / 1M / 5M for every (τ, K) in the sweep. Standard MV at the top for reference. Bold marks the best (τ, K, weight) per benchmark at each budget.

			GPT-OSS-20B											
			FrontierScience-Olympiad			HMMT Feb 2026			AIME 2025			Brumo 2025		
τ	K	Method	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M	B=250k	B=1M	B=5M
		Standard MV	$.482_{\pm.002}$	$.523_{\pm.001}$	$.537_{\pm.001}$	$.736_{\pm.003}$	$.775_{\pm.002}$	$.807_{\pm.001}$	$.881_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.901_{\pm.003}$	$.922_{\pm.002}$	$.924_{\pm.001}$
0.75	1	PC-linear	$.495_{\pm.002}$	$.524_{\pm.001}$	$.535_{\pm.001}$	$.751_{\pm.003}$	$.797_{\pm.002}$	$.812_{\pm.001}$	$.887_{\pm.002}$	$.899_{\pm.001}$	$.900_{\pm.000}$	$.903_{\pm.003}$	$.928_{\pm.002}$	$.934_{\pm.001}$
		PC-quadratic	$.502_{\pm.002}$	$.532_{\pm.001}$	$.537_{\pm.001}$	$.752_{\pm.003}$	$\mathbf{.800}_{\pm.002}$	$.814_{\pm.001}$	$\mathbf{.888}_{\pm.002}$	$.901_{\pm.002}$	$.901_{\pm.001}$	$.904_{\pm.003}$	$.930_{\pm.002}$	$.940_{\pm.001}$
		PC-cubic	$.502_{\pm.002}$	$.537_{\pm.001}$	$.545_{\pm.001}$	$\mathbf{.753}_{\pm.004}$	$.797_{\pm.002}$	$.814_{\pm.001}$	$.888_{\pm.002}$	$\mathbf{.902}_{\pm.002}$	$.902_{\pm.001}$	$.903_{\pm.003}$	$.931_{\pm.002}$	$\mathbf{.947}_{\pm.002}$
	2	PC-linear	$.495_{\pm.002}$	$.525_{\pm.001}$	$.531_{\pm.001}$	$.745_{\pm.004}$	$.795_{\pm.002}$	$.809_{\pm.001}$	$.883_{\pm.002}$	$.896_{\pm.002}$	$.900_{\pm.001}$	$.901_{\pm.003}$	$.927_{\pm.002}$	$.934_{\pm.001}$
		PC-quadratic	$.500_{\pm.002}$	$.532_{\pm.001}$	$.537_{\pm.001}$	$.745_{\pm.004}$	$.799_{\pm.002}$	$.815_{\pm.001}$	$.883_{\pm.003}$	$.899_{\pm.002}$	$.901_{\pm.001}$	$.900_{\pm.003}$	$.929_{\pm.002}$	$.942_{\pm.002}$
		PC-cubic	$.498_{\pm.002}$	$.530_{\pm.001}$	$.540_{\pm.001}$	$.745_{\pm.004}$	$.798_{\pm.003}$	$.817_{\pm.001}$	$.882_{\pm.003}$	$.901_{\pm.002}$	$.905_{\pm.001}$	$.900_{\pm.003}$	$.929_{\pm.002}$	$.946_{\pm.002}$
	3	PC-linear	$.500_{\pm.002}$	$.528_{\pm.001}$	$.534_{\pm.001}$	$.740_{\pm.004}$	$.793_{\pm.002}$	$.811_{\pm.001}$	$.881_{\pm.002}$	$.896_{\pm.002}$	$.900_{\pm.001}$	$.899_{\pm.003}$	$.926_{\pm.002}$	$.936_{\pm.001}$
		PC-quadratic	$\mathbf{.504}_{\pm.002}$	$.534_{\pm.001}$	$.541_{\pm.001}$	$.742_{\pm.004}$	$.800_{\pm.002}$	$.817_{\pm.001}$	$.881_{\pm.003}$	$.899_{\pm.002}$	$.903_{\pm.001}$	$.898_{\pm.003}$	$.929_{\pm.002}$	$.943_{\pm.001}$
		PC-cubic	$.501_{\pm.002}$	$.534_{\pm.001}$	$.544_{\pm.001}$	$.742_{\pm.004}$	$.799_{\pm.003}$	$\mathbf{.822}_{\pm.001}$	$.880_{\pm.003}$	$.900_{\pm.002}$	$\mathbf{.905}_{\pm.002}$	$.898_{\pm.003}$	$.928_{\pm.002}$	$.946_{\pm.002}$
0.50	1	PC-linear	$.489_{\pm.002}$	$.524_{\pm.001}$	$.540_{\pm.001}$	$.728_{\pm.003}$	$.773_{\pm.003}$	$.805_{\pm.002}$	$.886_{\pm.002}$	$.900_{\pm.001}$	$.900_{\pm.000}$	$\mathbf{.906}_{\pm.003}$	$.931_{\pm.002}$	$.935_{\pm.001}$
		PC-quadratic	$.494_{\pm.002}$	$.531_{\pm.001}$	$.541_{\pm.001}$	$.729_{\pm.003}$	$.779_{\pm.003}$	$.809_{\pm.001}$	$.885_{\pm.002}$	$.899_{\pm.001}$	$.900_{\pm.000}$	$.906_{\pm.003}$	$\mathbf{.932}_{\pm.002}$	$.938_{\pm.001}$
		PC-cubic	$.493_{\pm.002}$	$.532_{\pm.001}$	$.541_{\pm.001}$	$.729_{\pm.004}$	$.780_{\pm.003}$	$.810_{\pm.001}$	$.885_{\pm.002}$	$.898_{\pm.002}$	$.900_{\pm.001}$	$.904_{\pm.003}$	$.931_{\pm.002}$	$.939_{\pm.001}$
	2	PC-linear	$.492_{\pm.002}$	$.527_{\pm.001}$	$.535_{\pm.001}$	$.727_{\pm.003}$	$.776_{\pm.003}$	$.805_{\pm.002}$	$.883_{\pm.002}$	$.900_{\pm.001}$	$.900_{\pm.000}$	$.891_{\pm.003}$	$.924_{\pm.002}$	$.931_{\pm.001}$
		PC-quadratic	$.496_{\pm.002}$	$.533_{\pm.001}$	$.542_{\pm.001}$	$.727_{\pm.004}$	$.780_{\pm.003}$	$.809_{\pm.001}$	$.880_{\pm.003}$	$.900_{\pm.002}$	$.900_{\pm.000}$	$.888_{\pm.003}$	$.924_{\pm.002}$	$.933_{\pm.001}$
		PC-cubic	$.495_{\pm.002}$	$.532_{\pm.001}$	$.545_{\pm.001}$	$.726_{\pm.004}$	$.779_{\pm.003}$	$.807_{\pm.001}$	$.880_{\pm.003}$	$.898_{\pm.002}$	$.901_{\pm.001}$	$.885_{\pm.003}$	$.922_{\pm.002}$	$.935_{\pm.001}$
	3	PC-linear	$.493_{\pm.002}$	$.530_{\pm.001}$	$.541_{\pm.001}$	$.721_{\pm.004}$	$.773_{\pm.003}$	$.801_{\pm.002}$	$.881_{\pm.003}$	$.900_{\pm.001}$	$.900_{\pm.000}$	$.889_{\pm.003}$	$.924_{\pm.002}$	$.932_{\pm.001}$
		PC-quadratic	$.499_{\pm.002}$	$.535_{\pm.001}$	$.539_{\pm.001}$	$.720_{\pm.004}$	$.779_{\pm.003}$	$.807_{\pm.001}$	$.880_{\pm.003}$	$.899_{\pm.002}$	$.900_{\pm.001}$	$.884_{\pm.003}$	$.923_{\pm.002}$	$.935_{\pm.001}$
		PC-cubic	$.500_{\pm.002}$	$.537_{\pm.001}$	$.548_{\pm.001}$	$.719_{\pm.004}$	$.778_{\pm.003}$	$.806_{\pm.002}$	$.879_{\pm.003}$	$.895_{\pm.002}$	$.900_{\pm.001}$	$.883_{\pm.003}$	$.919_{\pm.002}$	$.937_{\pm.001}$
0.25	1	PC-linear	$.487_{\pm.002}$	$.527_{\pm.001}$	$.538_{\pm.001}$	$.721_{\pm.003}$	$.765_{\pm.003}$	$.798_{\pm.002}$	$.880_{\pm.002}$	$.896_{\pm.001}$	$.900_{\pm.000}$	$.897_{\pm.003}$	$.925_{\pm.002}$	$.933_{\pm.001}$
		PC-quadratic	$.494_{\pm.002}$	$.537_{\pm.001}$	$.547_{\pm.001}$	$.723_{\pm.004}$	$.773_{\pm.003}$	$.807_{\pm.002}$	$.879_{\pm.002}$	$.894_{\pm.001}$	$.899_{\pm.000}$	$.899_{\pm.003}$	$.927_{\pm.002}$	$.934_{\pm.001}$
		PC-cubic	$.496_{\pm.002}$	$.542_{\pm.001}$	$.551_{\pm.001}$	$.723_{\pm.004}$	$.774_{\pm.003}$	$.812_{\pm.001}$	$.879_{\pm.002}$	$.891_{\pm.002}$	$.897_{\pm.001}$	$.898_{\pm.003}$	$.927_{\pm.002}$	$.935_{\pm.001}$
	2	PC-linear	$.495_{\pm.002}$	$.538_{\pm.001}$	$.549_{\pm.001}$	$.714_{\pm.004}$	$.761_{\pm.003}$	$.790_{\pm.002}$	$.879_{\pm.002}$	$.894_{\pm.001}$	$.900_{\pm.000}$	$.897_{\pm.003}$	$.928_{\pm.002}$	$.933_{\pm.001}$
		PC-quadratic	$.500_{\pm.002}$	$.544_{\pm.001}$	$.550_{\pm.001}$	$.715_{\pm.004}$	$.762_{\pm.003}$	$.798_{\pm.002}$	$.877_{\pm.002}$	$.892_{\pm.001}$	$.899_{\pm.001}$	$.893_{\pm.003}$	$.927_{\pm.002}$	$.933_{\pm.001}$
		PC-cubic	$.500_{\pm.002}$	$\mathbf{.548}_{\pm.001}$	$.554_{\pm.001}$	$.716_{\pm.004}$	$.762_{\pm.003}$	$.802_{\pm.002}$	$.877_{\pm.002}$	$.888_{\pm.002}$	$.896_{\pm.001}$	$.891_{\pm.003}$	$.926_{\pm.002}$	$.933_{\pm.000}$
	3	PC-linear	$.488_{\pm.002}$	$.535_{\pm.001}$	$.549_{\pm.001}$	$.707_{\pm.004}$	$.761_{\pm.003}$	$.793_{\pm.002}$	$.876_{\pm.003}$	$.895_{\pm.001}$	$.900_{\pm.000}$	$.887_{\pm.003}$	$.924_{\pm.002}$	$.930_{\pm.001}$
		PC-quadratic	$.493_{\pm.002}$	$.544_{\pm.001}$	$.553_{\pm.001}$	$.708_{\pm.004}$	$.761_{\pm.003}$	$.795_{\pm.002}$	$.875_{\pm.003}$	$.893_{\pm.001}$	$.899_{\pm.001}$	$.883_{\pm.003}$	$.923_{\pm.002}$	$.931_{\pm.001}$
		PC-cubic	$.493_{\pm.002}$	$.547_{\pm.001}$	$\mathbf{.562}_{\pm.001}$	$.709_{\pm.004}$	$.761_{\pm.003}$	$.794_{\pm.002}$	$.874_{\pm.003}$	$.890_{\pm.002}$	$.896_{\pm.001}$	$.882_{\pm.003}$	$.921_{\pm.002}$	$.931_{\pm.001}$

Table 17 reports PC-linear, PC-quadratic, and PC-cubic accuracy across all (τ, K) configurations on the four GPT-OSS-20B benchmarks at budgets $B \in \{250\text{k}, 1\text{M}, 5\text{M}\}$, with Standard MV at the top for reference. The main paper fixes $\tau{=}0.75$, $K{=}1$ (Section 4) as a minimal default that preserves most of the trace and uses a single regeneration per sample. We use the sweep below to characterize how PC-WMV accuracy varies with (τ, K), not to select the operating point.

Two observations follow. At 250k tokens, high-cost configurations (large K and small τ) underperform Standard MV because each regeneration eats into the budget for new groups: $\tau{=}0.25$, $K{=}3$ PC-cubic (cost $\approx 3.25\times$) reaches .493, .709, .874, and .882 on FrontierScience-Olympiad, HMMT, AIME, and Brumo, underperforming Standard MV on the latter three (.736, .881, .901), while the

low-cost τ=0.75, K=1 (cost ≈1.25×) outperforms Standard MV on all four. At 1M–5M budgets, the default τ=0.75, K=1 stays close to the per-cell best on the three math benchmarks (HMMT, AIME, Brumo). FrontierScience-Olympiad is the exception, where some higher-cost configurations improve accuracy (PC-cubic at τ=0.25, K=3 reaches .547 at 1M and .562 at 5M on FrontierScience-Olympiad, against .537 and .545 at τ=0.75, K=1). K=1 is sufficient on the math benchmarks at 1M, while K=3 yields modest improvements on HMMT at 5M (PC-cubic .822 at K=3 vs. .814 at K=1, both at τ=0.75). Overall, this GPT-OSS-20B sweep supports the effectiveness of the default τ=0.75, K=1 on the four benchmarks, but does not establish insensitivity to (τ, K) for the other models.

D.6 Robustness to Judge Choice

We show that PC's per-cell advantage over the best baseline holds regardless of whether answer equivalence is evaluated by the model itself or by a stronger external grader.

The main paper uses each evaluated model itself to determine answer equivalence (Appendix H.2). One possible concern is that the evaluation may be biased by the limited capability of the model itself. As a robustness check, this section reports results using a stronger external grader. Namely, we re-scored the released pool with Claude Sonnet 4.6 [Anthropic, 2026] on four models (GPT-OSS-120B, GPT-OSS-20B, Nemotron3-30B, Nemotron2-9B) across the three LLM-judged benchmarks (Brumo 2025, HMMT Feb 2026, FrontierScience-Olympiad), giving 12 cells. AIME 2025 is exact-match and is therefore judge-free.

We re-cluster the released generation pool using the external judge's YES/NO equivalence judgments instead of the model's own, then recompute per-sample correctness labels along with $\overline{\text{AUROC}}$, PC-WMV accuracy, and the baseline accuracies. No new samples are drawn from the model.

PC's $\overline{\text{AUROC}}$ advantage is preserved on every cell. Figure 9 plots, per cell, the gap $\Delta\overline{\text{AUROC}} := \overline{\text{AUROC}}_{\text{PC}} - \overline{\text{AUROC}}_{\text{best baseline}}$ under self-judge (x) against the same quantity under the external judge (y). All 12 points fall in the same quadrant under both judges: 9 cells with $\Delta\overline{\text{AUROC}} > 0$ (PC ahead of the best baseline) and 3 cells with $\Delta\overline{\text{AUROC}} < 0$. No cell flips. Per-cell numbers are in Table 18.

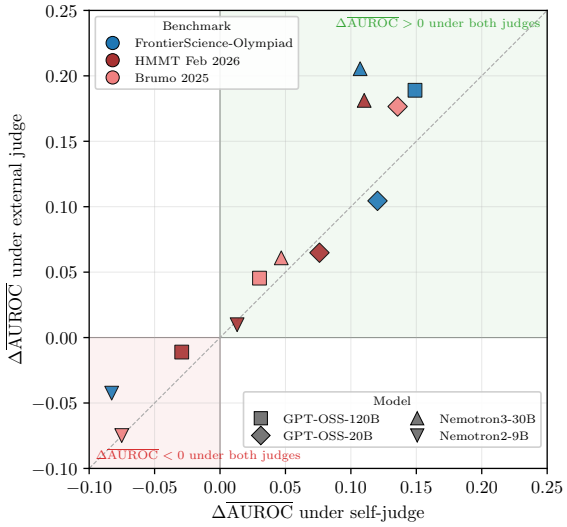


Figure 9: **PC's $\overline{\text{AUROC}}$ advantage is preserved across judges.** Each axis is $\Delta\overline{\text{AUROC}} := \overline{\text{AUROC}}_{\text{PC}} - \overline{\text{AUROC}}_{\text{best baseline}}$, under self-judge ($x$) and an external judge (Claude Sonnet 4.6, y); one point per cell (12 cells), and the dashed line is y=x. All 12 points lie in the top-right or bottom-left quadrant; no cell crosses an axis.

Table 18: **Per-cell PC and best-baseline $\overline{\text{AUROC}}$ under self-judge vs. an external judge (Claude Sonnet 4.6).** Bold marks the higher of PC and the best baseline within each cell, with the best baseline's identity in parentheses. AIME 2025 is exact-match (judge-free) and is omitted.

Model	Benchmark	PC		best baseline	
		self	external	self	external
GPT-OSS-120B	FrontierScience-Olympiad	**0.719**	**0.778**	0.570 (Self-certainty)	0.589 (Self-certainty)
	HMMT Feb 2026	0.698	0.711	**0.728** (DeepConf tail)	**0.722** (DeepConf tail)
	Brumo 2025	**0.636**	**0.660**	0.606 (DeepConf tail)	0.615 (DeepConf tail)
GPT-OSS-20B	FrontierScience-Olympiad	**0.707**	**0.718**	0.587 (P(True))	0.613 (P(True))
	HMMT Feb 2026	**0.743**	**0.740**	0.667 (P(True))	0.675 (P(True))
	Brumo 2025	**0.719**	**0.736**	0.583 (P(True))	0.559 (P(True))
Nemotron3-30B	FrontierScience-Olympiad	**0.631**	**0.733**	0.524 (Verbal 0–100)	0.527 (Verbal 0–100)
	HMMT Feb 2026	**0.698**	**0.766**	0.588 (DeepConf tail)	0.584 (P(True))
	Brumo 2025	**0.801**	**0.823**	0.755 (DeepConf tail)	0.762 (DeepConf tail)
Nemotron2-9B	FrontierScience-Olympiad	0.537	0.603	**0.620** (P(True))	**0.646** (P(True))
	HMMT Feb 2026	**0.648**	**0.648**	0.635 (P(True))	0.639 (P(True))
	Brumo 2025	0.633	0.633	**0.708** (P(True))	**0.708** (P(True))

WMV cost–accuracy curves under both judges nearly overlap. Figure 10 overlays the WMV cost–accuracy curves under self-judge (faded) and the external judge (saturated) for the four primary methods (PC-cubic, Standard MV, DeepConf tail, P(True)). On Brumo and HMMT, the two layers are essentially indistinguishable. On FrontierScience-Olympiad, the external-judge curves sit above their self-judge counterparts by a roughly uniform vertical offset; PC-cubic stays the top (or tied-top) curve in every panel where it leads under self-judge, and stays below where it trails. The curves show broadly similar cost–accuracy trade-offs under both judges in this tested subset.

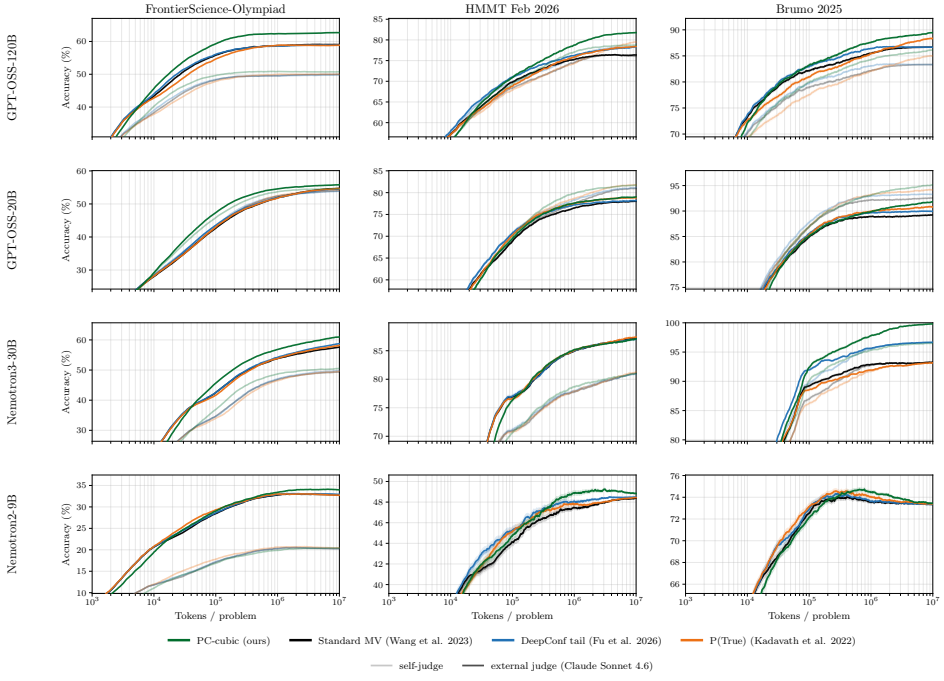


Figure 10: **WMV cost–accuracy curves are stable across judges.** 4 models × 3 benchmarks; faded curves use the model's own judge, saturated curves use an external judge (Claude Sonnet 4.6). The four methods (PC-cubic, Standard MV, DeepConf tail, P(True)) match Figure 5. The broad relative performance of the methods is similar across the panels; the largest absolute shifts between the two judges are in the FrontierScience-Olympiad column.

Pool-level shifts are concentrated on FrontierScience-Olympiad. On Brumo and HMMT, the two judges disagree on fewer than 4% of per-sample correctness labels per cell, and Pass@1 differs by at most 0.06. On FrontierScience-Olympiad the per-sample flip rate is higher (9% to 17%) and Pass@1 rises by up to 0.09 on three of the four cells; the fourth (GPT-OSS-20B) has compensating up- and down-flips that leave Pass@1 within 0.01 despite an 8.6% flip rate. The shift reflects the external judge recognizing more equivalent surface forms than the model's own. The added equivalences strengthen PC's relative position rather than weaken it: in Figure 9, three of the four FrontierScience-Olympiad points (blue) sit above the $y{=}x$ diagonal, and the leftmost column of Figure 10 shows PC's gap over the baselines widening on those cells under the external judge.

D.7 Reliability Signals vs. Pass@1

This subsection backs the analysis of Section 4.4 (rising r_C and a smaller, problem-dependent r_W slope within each category) with full panels and slope tables for the reproduction rates r_C, r_W under the GLM estimator (Appendix D.7.1), verifies that the qualitative pattern survives per-benchmark refits (Appendix D.7.2) and two alternative estimators (Appendix D.7.3), and reports the corresponding Pass@1 view for the per-sample baseline signals (Appendix D.7.4).

D.7.1 GLM Panels and Slopes

Figure 11 adds GLM panels for the two models not shown in Figure 3 (GPT-OSS-20B, Nemotron2-9B), and Table 19 reports the slopes for all five. The qualitative pattern is consistent: r_C rises with Pass@1 on both categories, while r_W has a smaller, problem-dependent slope and sits at a lower level on Science than on Math. The GLM fit kernel and cluster-bootstrap inference protocol below are shared with the alternative estimators of Appendix D.7.3.

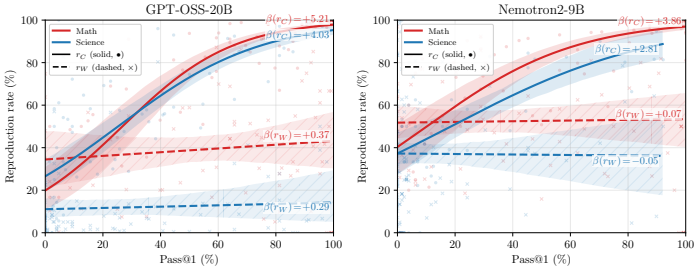


Figure 11: **Per-problem reproduction rates vs. Pass@1, remaining models.** GLM fits for GPT-OSS-20B and Nemotron2-9B, complementing Figure 3. Curves and overlays follow the same convention.

Table 19: **Logistic GLM slope estimates on** $\mathrm{logit}(r) = \beta_0 + \beta \cdot$ Pass@1 **per (model, category).** Bootstrap percentile CIs are from cluster bootstrap over problems (1000 replicates). The p-column gives the two-sided bootstrap p-value for $H_0 : \beta = 0$, with "<.001" indicating that no replicate crossed zero.

Model	Category	$\beta(r_C)$		$\beta(r_W)$	
		$\hat{\beta}$ [CI]	p	$\hat{\beta}$ [CI]	p
GPT-OSS-120B	Science	$+4.28$ $[+3.51, +5.14]$	<.001	-0.06 $[-1.20, +0.99]$	0.96
	Math	$+5.10$ $[+4.05, +6.26]$	<.001	$+1.14$ $[+0.04, +2.30]$	0.034
GPT-OSS-20B	Science	$+4.03$ $[+3.36, +4.77]$	<.001	$+0.29$ $[-1.05, +1.41]$	0.66
	Math	$+5.21$ $[+4.38, +6.25]$	<.001	$+0.37$ $[-0.67, +1.56]$	0.48
Nemotron3-30B	Science	$+3.67$ $[+2.90, +4.37]$	<.001	$+0.14$ $[-1.03, +1.20]$	0.8
	Math	$+3.18$ $[+2.17, +4.37]$	<.001	-0.20 $[-1.40, +0.79]$	0.67
Nemotron2-9B	Science	$+2.81$ $[+1.89, +4.07]$	<.001	-0.05 $[-1.25, +0.90]$	0.89
	Math	$+3.86$ $[+3.21, +4.69]$	<.001	$+0.07$ $[-0.69, +0.74]$	0.85
Ministral3-14B	Science	$+2.66$ $[+1.27, +4.09]$	0.002	-0.61 $[-1.98, +0.55]$	0.27
	Math	$+2.94$ $[+2.22, +3.67]$	<.001	-0.75 $[-1.45, -0.08]$	0.022

GLM fit kernel. For each outcome $Y \in \{C, W\}$ (regenerations seeded from a correct (C) or wrong (W) trace, matching r_C and r_W in Section 3.1), fit a logistic regression $\Pr(Y_j = 1 \mid \text{Pass@1}_j) = \sigma(\beta_{0,Y} + \beta_Y \cdot \text{Pass@1}_j)$ on trial-level data, where each trial j inherits the Pass@1 of its problem, by maximum likelihood (`statsmodels.GLM`, Binomial family, default iteratively reweighted least squares solver). Reported quantities are the slope β_Y and the predicted curve on a fixed Pass@1 grid, masked to the cell's observed support.

Inference protocol. A (model, category) cell consists of N per-problem records, each carrying its Pass@1 and two Bernoulli trial streams: $r_{C,i,j} \in \{0, 1\}$ for the $n_{C,i}$ regenerations seeded from a correct trace and $r_{W,i,j}$ for the $n_{W,i}$ from a wrong trace ($r = 1$ if the regenerated answer matches the seed, 0 otherwise). Each problem i contributes $n_{C,i}$ trials $\big(\text{Pass@1}_i,\, r_{C,i,j}\big)$ to the r_C regression (and $n_{W,i}$ trials $\big(\text{Pass@1}_i,\, r_{W,i,j}\big)$ to r_W), where the same Pass@1 is shared across all trials from the same problem. We refer to this trial-level dataset as the expanded trials. Point estimates apply each estimator's fit kernel $\hat{f}_Y$ ($Y \in \{C, W\}$) to the expanded trials.

Confidence intervals are derived using $B = 1000$ cluster-bootstrap iterations. At iteration b, we resample N problem indices with replacement, then refit $\hat{f}_C^{(b)}$ and $\hat{f}_W^{(b)}$ on the corresponding trial set. Within-problem correlation among regenerations that share a prefix is corrected for by resampling problems rather than trials; trial-level Bernoulli standard errors would otherwise be underestimated by a factor of 2 to 4 on these data. Because $\hat{f}_C^{(b)}$ and $\hat{f}_W^{(b)}$ share the same resample within an iteration, the difference $D^{(b)} = \hat{f}_C^{(b)} - \hat{f}_W^{(b)}$ inherits the correct joint distribution, and CIs for D are read off $\{D^{(b)}\}$ directly rather than summed from per-fit half-widths.

We report 2σ-equivalent bootstrap percentile CIs: their endpoints are the $\Phi(-2)$ and $\Phi(2)$ quantiles of the bootstrap distribution, matching the central probability mass within two standard deviations of a normal distribution. This specifies the nominal coverage, not a point estimate ± 2 bootstrap standard deviations. Curve intervals are pointwise, taken per Pass@1 grid point for continuous estimators and per bin for the binned estimator. The two-sided bootstrap p-value for $H_0 : \beta = 0$ in Table 19 is $p = \min\{1,\, 2\min(\Pr_b[\beta^{(b)} \le 0],\, \Pr_b[\beta^{(b)} \ge 0])\}$, with "<.001" indicating no replicate crossed zero. The seed is fixed at 0. A small convergence check (3 seeds, several B, GPT-OSS-120B Math r_W) confirmed stability from $B = 1000$.

D.7.2 Robustness under Per-Benchmark Pooling

Section 4.4 pools the three Math benchmarks (HMMT Feb 2026, AIME 2025, Brumo 2025) into a single Math curve per model. Since the three span different Pass@1 ranges within a fixed model, the pooled slope could in principle reflect between-benchmark Pass@1 differences rather than a within-benchmark Pass@1 effect.

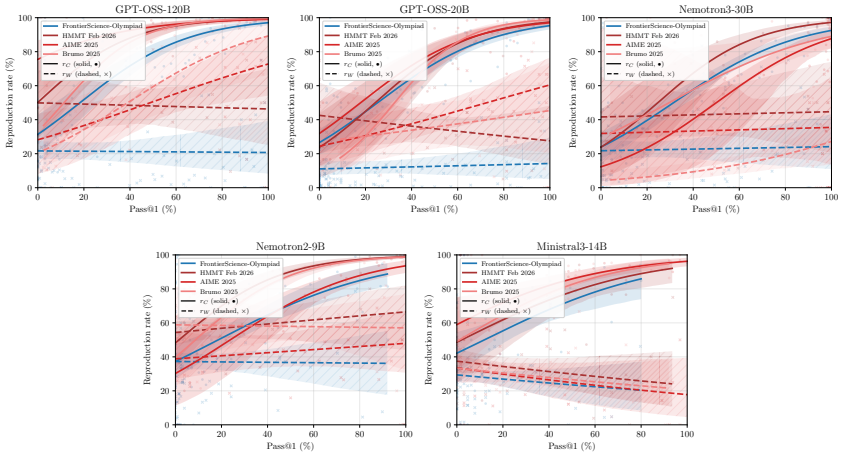


Figure 12: **Per-benchmark reproduction rates vs. Pass@1.** One panel per model. Solid (•): r_C. Dashed (×): r_W. Lines are logistic GLM fits with pointwise cluster-bootstrap CIs. Scatter shows per-problem rates. Unlike Figure 3, the Math benchmarks are fitted separately.

For the per-benchmark analysis, we refit the same logistic GLM on each (model, benchmark) cell separately, with the fit kernel and cluster-bootstrap protocol of Appendix D.7.1. Figure 12 plots the four per-benchmark curves per model in a single panel, and Table 20 reports the slopes.

Table 20: **Per-(model, benchmark) GLM slope estimates on** $\text{logit}(r) = \beta_0 + \beta \cdot \text{Pass@1}$**.** Per-benchmark variant of Table 19, fit on each (model, benchmark) cell separately. Bootstrap percentile CIs from a cluster bootstrap over problems (1000 replicates), and two-sided bootstrap p-values for $H_0 : \beta = 0$ ("<.001" indicates that no replicate crossed zero).

Model	Benchmark	$\beta(r_C)$		$\beta(r_W)$	
		$\hat{\beta}$ [CI]	p	$\hat{\beta}$ [CI]	p
GPT-OSS-120B	FrontierScience-Olympiad	$+4.28$ $[+3.51, +5.14]$	<.001	-0.06 $[-1.20, +0.99]$	0.96
	HMMT Feb 2026	$+4.97$ $[+3.76, +6.45]$	<.001	-0.14 $[-1.58, +1.65]$	0.83
	AIME 2025	$+3.50$ $[+2.01, +5.80]$	<.001	$+1.92$ $[+0.67, +4.50]$	0.002
	Brumo 2025	$+6.21$ $[+4.53, +8.34]$	<.001	$+3.46$ $[+1.25, +5.75]$	0.004
GPT-OSS-20B	FrontierScience-Olympiad	$+4.03$ $[+3.36, +4.77]$	<.001	$+0.29$ $[-1.05, +1.41]$	0.66
	HMMT Feb 2026	$+4.82$ $[+3.61, +6.83]$	<.001	-0.66 $[-2.01, +1.17]$	0.4
	AIME 2025	$+4.24$ $[+3.05, +6.08]$	<.001	$+1.55$ $[+0.45, +2.91]$	0.006
	Brumo 2025	$+6.55$ $[+5.28, +8.25]$	<.001	$+0.81$ $[-1.34, +3.16]$	0.42
Nemotron3-30B	FrontierScience-Olympiad	$+3.67$ $[+2.90, +4.37]$	<.001	$+0.14$ $[-1.03, +1.20]$	0.8
	HMMT Feb 2026	$+4.72$ $[+3.48, +6.71]$	<.001	$+0.13$ $[-1.75, +1.68]$	0.85
	AIME 2025	$+3.93$ $[+2.59, +6.97]$	<.001	$+0.16$ $[-4.10, +3.47]$	0.86
	Brumo 2025	$+2.97$ $[+0.40, +5.26]$	0.016	$+2.14$ $[+0.68, +3.59]$	0.018
Nemotron2-9B	FrontierScience-Olympiad	$+2.81$ $[+1.89, +4.07]$	<.001	-0.05 $[-1.25, +0.90]$	0.89
	HMMT Feb 2026	$+4.57$ $[+3.43, +5.93]$	<.001	$+0.51$ $[-0.49, +1.45]$	0.26
	AIME 2025	$+3.52$ $[+2.30, +5.31]$	<.001	$+0.37$ $[-0.74, +1.40]$	0.42
	Brumo 2025	$+4.83$ $[+3.96, +6.00]$	<.001	-0.07 $[-2.18, +1.47]$	1
Ministral3-14B	FrontierScience-Olympiad	$+2.66$ $[+1.27, +4.09]$	0.002	-0.61 $[-1.98, +0.55]$	0.27
	HMMT Feb 2026	$+2.68$ $[+0.87, +4.48]$	0.006	-0.68 $[-1.84, +0.49]$	0.2
	AIME 2025	$+2.88$ $[+1.71, +4.13]$	<.001	-0.85 $[-2.44, +0.42]$	0.22
	Brumo 2025	$+3.19$ $[+2.01, +4.98]$	<.001	-0.65 $[-1.69, +0.56]$	0.28

The two claims in Section 4.4 both survive the per-benchmark refit. (i) $\beta(r_C) > 0$ on all 20 (model, benchmark) cells with $p < 0.05$ (in fact $p < 0.001$ on 17 of 20 cells), so the rising-r_C pattern is within-benchmark and not a pooling artefact. (ii) $|\beta(r_W)| < \beta(r_C)$ on all 20 cells, so the smaller-magnitude statement is also within-benchmark.

The per-benchmark view further clarifies the two non-zero pooled $\beta(r_W)$ values from Section 4.4. The GPT-OSS-120B Math value $+1.14$ is driven by AIME 2025 ($+1.92$, $p = 0.002$) and Brumo 2025 ($+3.46$, $p = 0.004$), with HMMT Feb 2026 separately null (-0.14, $p = 0.83$): the pooled value reflects a real within-benchmark Pass@1 effect on AIME and Brumo, but the effect is benchmark-specific rather than uniform across Math. The Ministral3-14B Math value -0.75 is uniform across all three benchmarks (-0.85, -0.68, -0.65), i.e. a within-benchmark effect that holds Math-wide. In neither case is the pooled slope a between-benchmark artefact.

D.7.3 Robustness under Alternative Estimators

In Figure 3, we fit a logit-linear model. To verify that the conclusions in the main paper (that is, rising r_C, a smaller and problem-dependent r_W slope within each category, and lower r_W on Science than on Math) do not depend on that form, we rerun the same plot for all five models under two alternative estimators, sharing the cluster-bootstrap protocol of Appendix D.7.1.

Trial-pooled binned estimator (Figure 13). This estimator reads reproduction rates directly off the data, removing the GLM's logit-linear assumption at the cost of bin-edge sensitivity. Bin edges come from the full-data Pass@1 quantiles, with 5 equal-count bins per category deduplicated by `np.unique`. A Pass@1 mass at zero can collapse adjacent edges, e.g. Ministral3-14B Science with 53 problems at zero resolves to 3 effective bins. Each bootstrap iteration assigns problems to those fixed edges and reports the trial-pooled rate $\sum_{i \in \text{bin}} k_{Y,i} / \sum_{i \in \text{bin}} n_{Y,i}$, where $k_{Y,i}$ and $n_{Y,i}$ count reproductions and trials.

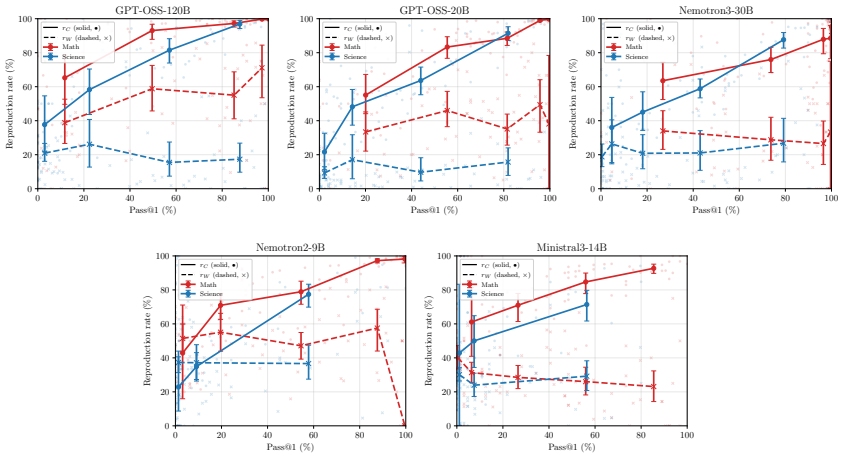


Figure 13: **Binned view of Figure 3, all five models.** Trial-pooled r_C (solid, •) and r_W (dashed, $\times$) over five Pass@1 quantile bins per category, with cluster-bootstrap CIs (1000 replicates, same inference protocol as the GLM in the main paper).

LOWESS smoother (Figure 14). Because this estimator does not assume a monotone or parametric relationship, it can reveal any non-monotonic structure in r_W that a logit-linear model would inadvertently hide.

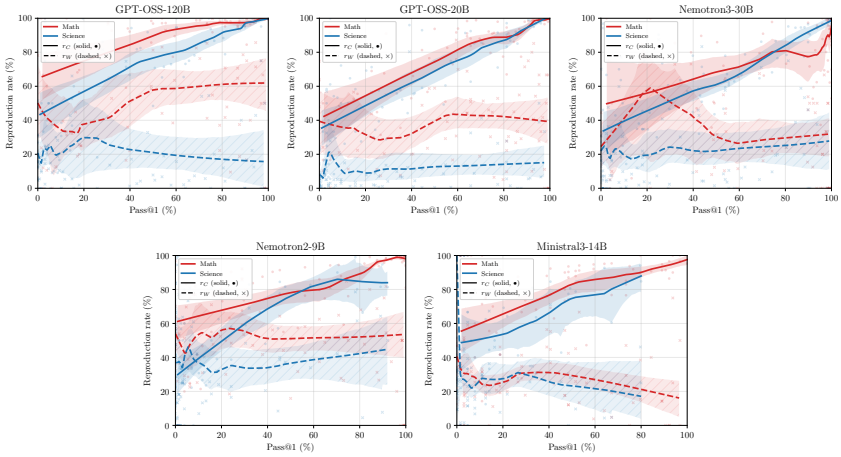


Figure 14: **LOWESS smoother view of Figure 3, all five models.** r_C (solid) and r_W (dashed) from trial-level LOWESS (span $= 0.5$, no robustness reweightings) per category, with pointwise bootstrap CIs from cluster bootstrap over problems (1000 replicates).

For the LOWESS fits, we use `statsmodels.nonparametric.smoothers_lowess` on trial-level $(\text{Pass@1}_j, Y_j)$ pairs with span $\text{frac} = 0.5$ and robustness iterations `it` $= 0$. The default `it` $= 3$ uses bisquare residual weights designed for continuous residuals, which collapse Bernoulli outcomes to degenerate 0 or 100% curves (verified empirically on Science). The fit is linearly interpolated onto the GLM's Pass@1 grid and masked to the observed support.

All three estimators agree qualitatively across all five models: r_C rises across Pass@1 on every (model, category) pair, and r_W has a smaller, problem-dependent slope with the category-specific levels described in Section 4.4.

D.7.4 Baseline Confidence Signals vs. Pass@1

For comparison with Figure 3, we repeat the per-class, Pass@1-conditioned view on the two baseline confidence signals named in the caption of Figure 2: DeepConf tail (Figure 15) and P(True)

(Figure 16). For each signal, we split per-trial values by whether the trace's answer matches gold (Correct, solid) or not (Wrong, dashed), and fit each class per category with a Gaussian linear model $s = \beta_0 + \beta \cdot \text{Pass@1}$, replacing the Bernoulli logistic GLM of Appendix D.7.1 since the response is now a continuous score. The cluster-bootstrap protocol ($B = 1000$ resamples over problems, 2σ percentile CIs) is unchanged. Scatter overlays show per-problem mean confidence (one dot per problem per class), the continuous-signal analogue of the per-problem rate k/n in Figure 3.

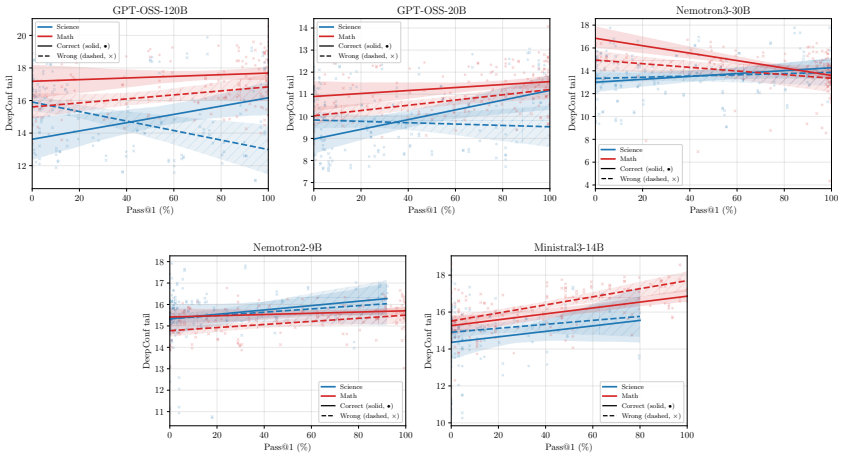


Figure 15: **DeepConf tail vs. Pass@1, all five models.** Per-class Gaussian linear fits with cluster-bootstrap CIs over problems. Scatter overlays are per-problem mean confidence.

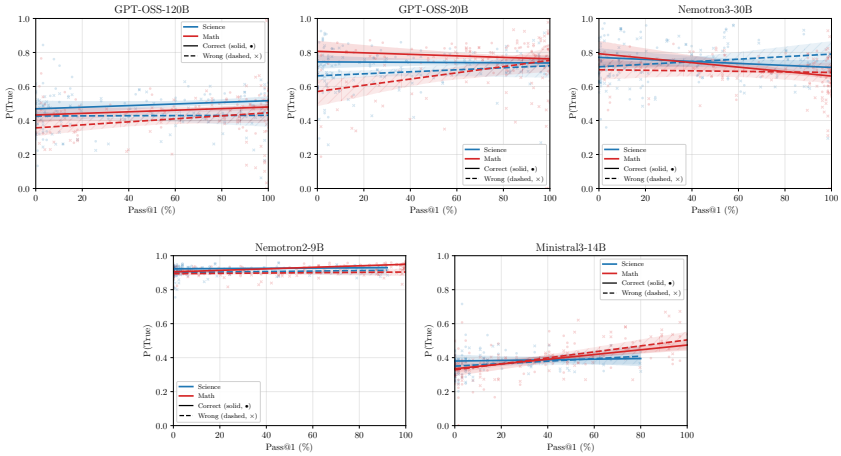


Figure 16: **P(True) vs. Pass@1, all five models.** Same protocol as Figure 15.

E Baseline Implementation Details

We document here the exact confidence scores and voting rules used for each baseline, since different papers use slightly different weighting conventions.

Glossary of baseline labels. The baselines summarized in Section 4 appear under the following labels in our tables and figures (e.g. Table 3 and the per-model tables of Appendix D.3). The per-baseline paragraphs below give the full implementation details.

- *DeepConf* [Fu et al., 2026]: a family of trace-level log-probability scores (*first-token*, *Mean* (the *Self-certainty* signal of Kang et al. [2025]), *bottom-10%*, *block-min*, *tail*). Each is plugged into Eq. (3) with w as the identity. Filtered variants (*top-10%*, *top-90%*) keep only the top-$\eta\%$ of traces by the same score before voting.

- *CISC* [Taubenfeld et al., 2025]: Confidence-Informed Self-Consistency, a per-sample weighting scheme that draws its raw confidence from one of four sources: *Response probability* [Wang et al., 2023b] (length-normalized geometric mean of per-token probabilities), *Verbal binary* [Lin et al., 2022] (a 0/1 self-rating parsed from a follow-up call), *Verbal 0–100* [Lin et al., 2022] (a percentage self-rating parsed from a follow-up call), and *P(True)* [Kadavath et al., 2022] (the renormalized softmax probability of the "1" token versus "0" at the rating-call confidence position).
- *Adaptive stopping*: cost–accuracy curves traced by sweeping the early-stopping hyperparameter of Adaptive Consistency (*AC sweep* [Aggarwal et al., 2023]) and Early-Stopping Self-Consistency (*ESC sweep* [Li et al., 2024]) over the same initial pool.
- *SubthoughtReasoner* [Hammoud et al., 2025]: a single-trace refinement baseline that segments each trace at linguistic cues and votes over per-subthought regenerations. It appears only in the GPT-OSS-20B all-baselines tables (Tables 7 and 12).
- *PC-linear, PC-quadratic, PC-cubic*: PC-WMV (Algorithm 1) instantiated with the power-family weighting $w^{(n)}(c) = c^n$ at $n = 1, 2, 3$. We use "PC-cubic" as shorthand for PC-WMV with n=3.

Deep Think with Confidence (DeepConf) [Fu et al., 2026]. We compute five trace-level confidence scores from the top-20 token log-probabilities saved at generation, following the original implementation.[3] For a trace y_i, let $P_{t,j}$ be the j-th largest token probability at position $t \le |y_i|$. The readout $\ell_i = \{P_{t,j}\}_{t \le |y_i|,\, j \le 20}$ is the top-20 log-probability record along y_i. Write $C_t = -\frac{1}{20}\sum_{j=1}^{20} \log P_{t,j}$ for the negative top-20 mean at position t. The five trace-level signals $s(y_i, \ell_i)$ are:

Score	Definition
First-token	$D_{\mathrm{KL}}(\widehat{P}_1 \| U)$, where $\widehat{P}_1$ renormalizes the available finite top-20 probabilities at the first generated token, and U is uniform on that support.
Mean	$\frac{1}{\lvert y_i \rvert}\sum_t C_t$, the average token-level confidence over the trace.
Bottom-10%	Mean of the lowest $\lceil 0.1m \rceil$ of m stride-1 moving averages over C_t. The window length is 1024, reduced to $\lfloor \lvert y_i \rvert / 2 \rfloor$ for shorter traces.
Block-min	Minimum block-level mean C_t, after splitting the trace at double-newline boundaries.
Tail	Mean C_t over the last $\min(\lvert y_i \rvert, 2024)$ tokens. The value 2024 is the implementation's literal constant.

The Mean score above is a top-20 implementation of the self-certainty signal originally proposed by Kang et al. [2025] on the full vocabulary. We adopt it as our *Self-certainty* baseline and report it under that label throughout the paper. DeepConf is the case of Eq. (3) where $s(y_i, \ell_i)$ is one of the five signals above and w is the identity, giving $v_i(a) = s(y_i, \ell_i) \cdot \mathbf{1}[a_i = a]$, where a_i is the answer extracted from trace i. Confidence filtering (Section 3.2 of their paper) optionally restricts the sum to the top-$\eta\%$ of traces by $s(y_i, \ell_i)$ before aggregation. The unfiltered variant (`DeepConf-<score>`) sums $v_i(a)$ over all traces, for each of the five scores above. In addition, we evaluate four filtered variants that combine the two stronger scores with the two retention levels proposed in the paper:

Score	η	Aggregation
Bottom-10%	10%	Keep top-10% by bottom-10% confidence, then sum $s(y_i, \ell_i) \cdot \mathbf{1}[a_i = a]$ over retained.
Bottom-10%	90%	Same with $\eta = 90\%$.
Tail	10%	Keep top-10% of traces by tail confidence, then sum $s(y_i, \ell_i) \cdot \mathbf{1}[a_i = a]$ over retained.
Tail	90%	Same with $\eta = 90\%$.

The DeepConf online early-termination mechanism is not used: our evaluation is offline over a pre-generated pool under a shared token budget.

Confidence-Informed Self-Consistency (CISC) [Taubenfeld et al., 2025]. CISC (Definition 3.1 of their paper) extracts a per-sample raw confidence $s_i = s(y_i, \ell_i)$ from one of four sources drawn from

[3] `https://github.com/facebookresearch/deepconf`

prior work, then applies a softmax weighting $w(s_i) = \exp(s_i/T)/\sum_j \exp(s_j/T)$ (over samples) with tunable temperature T:

Source	ℓ_i	Raw confidence s_i
Response probability [Wang et al., 2023b]	per-token log-probabilities of y_i	Length-normalized geometric mean of token probabilities, $\exp\big(\frac{1}{\lvert y_i\rvert}\sum_t \log P_t\big)$ where P_t is the probability of the generated token at position t.
Verbal binary [Lin et al., 2022]	$\emptyset$ (verbalized)	$\{0, 1\}$ self-rating parsed from text, a simplification of the verbalized paradigm.
Verbal 0–100 [Lin et al., 2022]	$\emptyset$ (verbalized)	Post-hoc percentage-confidence self-rating parsed from text.
P(True) [Kadavath et al., 2022]	top-20 logprobs at the rating-call confidence token	Renormalized softmax probability of "1" versus "0" computed from the top-20 logprobs at the rating-call confidence token, $\exp(\ell_1)/(\exp(\ell_0) + \exp(\ell_1))$ when both are present (with single-side fallbacks otherwise).

The original CISC release[4] targets non-reasoning models and forces a single-token response (`temperature=0`, `max_tokens=1`). Our reasoning models may emit channel-control tokens or reasoning before the confidence digit because the saved trace leaves the final channel open (e.g., Harmony for GPT-OSS). We therefore allow a separate completion from each saved trace y_i, using the model's recommended sampling parameters.

Verbal binary and P(True) share the suffix "*Now I will rate my confidence in the proposed answer as either 0 or 1. Proposed confidence: (*", while Verbal 0–100 uses "*Now I will rate my confidence in the proposed answer on a scale of 0–100. Proposed confidence: (*" (with leading newline). The wording mirrors the CISC release; one binary call supplies both Verbal binary and P(True). Appendix H.1 gives the extraction and missing-score rules. To avoid a baseline-specific T search that could over-fit our benchmarks, we report the untuned linear weighting $w(s) = s$ for each source (verbal 0–100 rescaled to $[0, 1]$), so $v_i(a) = s(y_i, \ell_i) \cdot \mathbf{1}[a_i = a]$. Softmax with T=1 is also computed by our pipeline but omitted from the tables because it consistently underperforms the linear variant. A full T sweep is left to future work.

Verbal baselines are charged $|y_i|$ plus the stored minimum-extraction token count, rather than the full secondary-call output (Appendix H.1). This favors the verbal baselines and is not their full incurred cost. The secondary call's maximum output length matches the initial generation budget (Appendix I). Response probability has no secondary call and reads the saved log-probabilities for free. PC-cubic sharpens its signal, while these confidence baselines use untuned linear weights; the AUROC comparison is invariant to strictly increasing score transformations.

Adaptive Consistency (AC) [Aggarwal et al., 2023]. AC consumes the initial pool one sample at a time and stops at the first k for which a Beta-binomial posterior favors the running top answer over the runner-up by at least C_{thresh}. Let n_1 and n_2 be the top-1 and top-2 answer counts after k samples. The closed-form stopping criterion is

$$1 - I_{1/2}(n_1 + 1,\, n_2 + 1) \;\geq\; C_{\text{thresh}},$$

where $I_x(\alpha, \beta)$ is the regularized incomplete beta function. This is the analytic equivalent of the Monte Carlo integral implemented in the official release.[5] At the stop time, the prediction is the running mode, so AC adds no per-sample weighting: it is Standard MV truncated at a natural stopping point. The original paper reports $C_{\text{thresh}} = 0.95$ as the default and sweeps $[0.5, 1)$ in Figure 2 to trace a cost–accuracy frontier. We sweep $C_{\text{thresh}} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99, 0.995, 0.999\}$. Each value yields a single (cost, accuracy) point at its natural stopping cost, and we label the resulting curve "AC sweep".

Early-Stopping Self-Consistency (ESC) [Li et al., 2024]. ESC consumes the initial pool in fixed-size windows of W samples. When a window is unanimous, ESC locks that answer as the final prediction and stops. Otherwise, the window's votes are added to a running counter and a new window is drawn. Before any lock, the intermediate prediction is the mode of the running counter.

[4] `https://github.com/google-research/google-research/tree/master/cisc`

[5] `https://github.com/Pranjal2041/AdaptiveConsistency`

The paper recommends $W = 5$ for most tasks and $W = 8$ for MATH. Our streaming reformulation reproduces the lock semantics of the official batch release.[6] We sweep $W \in \{2, 3, \ldots, 10\}$ and label the resulting natural-stopping curve "ESC sweep". Like AC, ESC is Standard MV with a window-based stopping rule and adds no per-sample weight.

Cost accounting for AC and ESC. Both methods consume only initial samples. Their per-trial budget is the cumulative generated-token length of the consumed initial samples, the same cost accounting used by Standard MV, Self-certainty, DeepConf, and Response probability. Because each hyperparameter value yields one (cost, accuracy) point at its natural stopping cost rather than a curve over a shared budget grid, we treat the points across the swept hyperparameter as the method's cost–accuracy curve and interpolate it for any required budget. Sampling is with replacement, so N does not cap the number of consumed samples or bound accuracy by N-sample Standard MV. The sampled sequence length is the larger of the maximum evaluated sample count and the maximum token budget divided by the cheapest pool sample's token cost, rounded down. If no stop occurs, the natural-stopping evaluation uses the vote and cost at the end of that finite sequence.

SubthoughtReasoner [Hammoud et al., 2025]. At the time of writing, no official implementation was available, so we reimplemented the method based on the description provided in the original paper. Our implementation segments each reasoning trace into sequential subthoughts at the linguistic cues described in Section 3 of their paper, regenerates a continuation from the end of each subthought, and takes an unweighted majority vote over the resulting pool of answers. To bound the per-problem cost of constructing this pool, we process initial samples in order, accumulating each sample's initial tokens together with its subthought-regenerated continuations, and stop once the cumulative cost reaches that of the N initial samples for that problem. Otherwise, the per-problem cost of regenerating subthought continuations from all N initial samples would be several times that of the initial generation alone. We evaluated this baseline on only a small subset of (model, benchmark) conditions.

Prefix consistency (ours). For prefix consistency, $\ell_i = \emptyset$: it reads only generated tokens and regenerated answers, with no log-probability access. Unlike the per-sample signal $s(y_i, \ell_i)$ used by the baselines above, our signal $c_i^{(\tau)}(a)$ is defined per distinct candidate $a \in A_i^{(\tau)}$ (Section 3.1). Moreover, we use the power-family weighting $w^{(n)}(c) = c^n$ for $n \in \{1, 2, 3\}$ (PC-linear, PC-quadratic, PC-cubic). PC-linear (n=1) admits a simple interpretation: substituting $w(c) = c$ and Eq. (7) into the PC-WMV vote gives $\sum_i v_i(a) = \frac{1}{K+1} \sum_i |\{a' \in A_i^{(\tau)} : a' = a\}|$, which is proportional to the count of a in the combined pool of N initial answers and NK regenerations. PC-linear's argmax therefore coincides with unweighted majority voting on this $(K+1)N$ pool, treating each regeneration as one additional vote of equal weight. PC-quadratic and PC-cubic depart from this baseline by giving super-linear weight to candidates that the same group reproduces.

F Examples of Prefix Consistency

Figure 17 shows two representative examples of how regeneration behaves after truncation.

When the initial answer is correct, the regenerated continuation often recovers the same core reasoning structure and reproduces the correct answer. In contrast, when the initial reasoning is already flawed, regeneration typically does not repair it. Instead, it continues along a similar erroneous line of reasoning and may even produce a different wrong answer. These examples help illustrate why regeneration amplifies the consistency of correct traces while still failing to escape wrong ones.

[6] `https://github.com/Yiwei98/ESC`

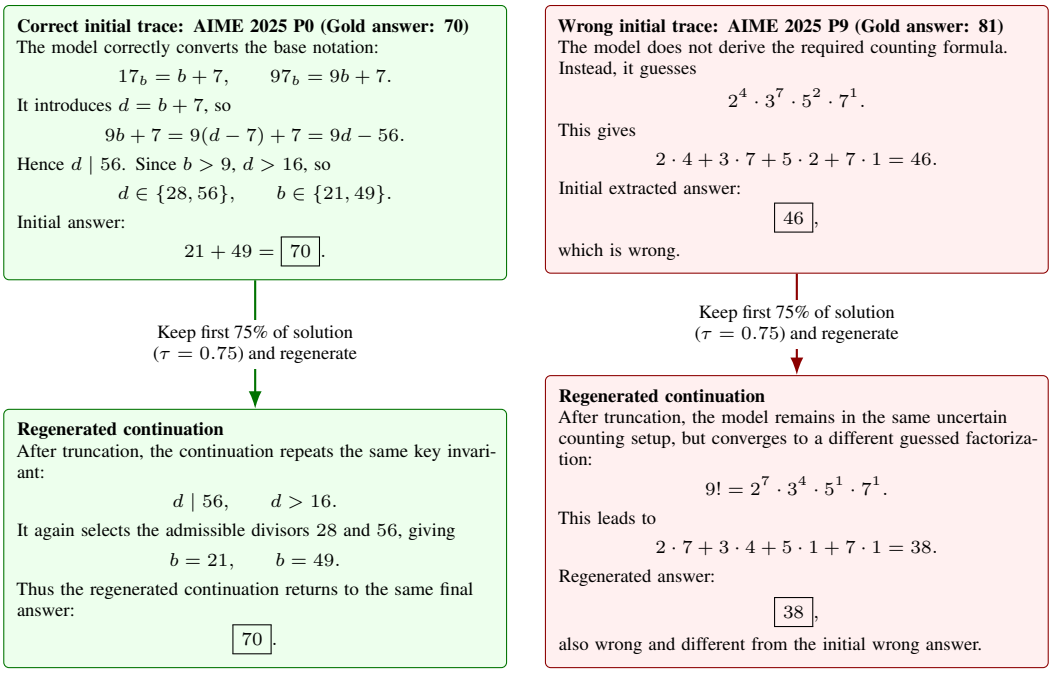


Figure 17: **Qualitative examples of CoT regeneration after truncation.** A correct initial trace (green, left) and a wrong initial trace (red, right) on AIME 2025. For each trace, the first 75% ($\tau = 0.75$) of the full generated solution (by token count, including the final answer) is retained and the remainder is regenerated.

G Evaluation Protocol

This appendix consolidates the analysis pipeline shared across the experimental sections of the main paper. Implementation choices specific to each baseline are in Appendix E. Hardware and inference settings are in Appendix I.

G.1 Setup

Benchmarks. We evaluate on one science benchmark (FrontierScience-Olympiad) and three math benchmarks (HMMT Feb 2026, AIME 2025, Brumo 2025). Table 21 lists the problem count, answer format, and scoring rule of each. We use the full released test split for every benchmark. Answers are extracted from the boxed expression, then math-normalized for the math benchmarks (fraction unification, degree-marker removal) or used with minimal normalization for FrontierScience-Olympiad. Equivalence to the gold answer is decided by exact match for AIME 2025 and by an LLM judge for the other three benchmarks (Appendix H.2). HuggingFace URLs and licenses are in Table 24.

Table 21: **Evaluation benchmarks.**

Dataset	# Problems	Category	Answer format	Scoring
FrontierScience-Olympiad	100	Science	Free-form (expression, number, or short text)	LLM judge
HMMT Feb 2026	33	Math	Integer or closed-form expression	LLM judge
AIME 2025	30	Math	Integer in $[0, 999]$	Exact match
Brumo 2025	30	Math	Integer or closed-form expression	LLM judge

Pre-generated pool and cost vectors. For each (model, benchmark) cell we draw N=128 initial generations per problem (N=64 for Ministral3-14B, where the full N=128 run had not completed at the time of writing), and for prefix-consistency methods we additionally draw K regenerations per initial sample at the chosen truncation fraction τ. All voting and analysis read from this fixed pool, so the same generated tokens back every comparison and the only randomness in the reported numbers

is from the analysis-side resampling described below. Each generation carries its actual recorded token count. For each method, the cost of a vote is the cumulative generated-token length of every sample it actually reads. Standard MV, Self-certainty, DeepConf, Response probability, AC, and ESC consume only the initial samples. CISC's verbalized sources (Verbal binary, Verbal 0–100) and P(True) additionally consume the secondary verbal-rating completions call per sample described in Appendix E. PC additionally reads the regenerated continuations from each sample's prefix, charging only newly generated continuation tokens, not the kept prefix again. SubthoughtReasoner reads the continuations regenerated from each subthought boundary.

Hyperparameter defaults. Unless otherwise noted, every result in the main paper uses τ=0.75, K=1, and the power-family weighting $w^{(n)}(c) = c^n$ for $n \in \{1, 2, 3\}$. Sensitivity to alternative choices of τ and K is reported in Appendix D.5.

Confidence intervals. For accuracy and mean cost, error bars show $\pm 2\times$ the Monte Carlo standard error. Token-efficiency ratio intervals are the point estimate ± 2 standard deviations of the parametric-bootstrap ratio distribution. Regression coefficients and fitted curves use 2σ-equivalent percentile intervals from a cluster bootstrap over problems. Their endpoints are the $\Phi(-2)$ and $\Phi(2)$ bootstrap quantiles, matching the nominal coverage of a normal 2σ interval. The trial Monte Carlo (Sections 4.2 and 4.3) draws $M = 500$ replicates per (method, benchmark, budget) cell. The parametric ratio bootstrap (Section 4.3) and the cluster bootstrap over problems (Section 4.4) each use 1,000 resamples.

G.2 $\overline{\mathrm{AUROC}}$ for the Correctness-Predictor Evaluation (Section 4.1)

Definition of $\overline{\mathrm{AUROC}}$. AUROC denotes the Area Under the Receiver Operating Characteristic curve. Let $\mathcal{Q}' := \{q \in \mathcal{Q} : 0 < \mathrm{Pass@1}_q < 1\}$ be the subset of problems on which AUROC_q and the per-problem rates $r_{C,q}, r_{W,q}$ are defined, and set $s_i = c_i(a_i)$ for prefix consistency and $s_i = s(y_i, \ell_i)$ for the baselines (Appendix E). We score each signal by a per-problem AUROC and macro-average over $\mathcal{Q}'$:

$$\begin{aligned} \mathrm{AUROC}_q &:= \Pr\big[s_i > s_j \,\big|\, a_i = a_q^\star,\, a_j \neq a_q^\star\big] \\ &\quad + \tfrac{1}{2}\Pr\big[s_i = s_j \,\big|\, a_i = a_q^\star,\, a_j \neq a_q^\star\big]\,, \\ \overline{\mathrm{AUROC}} &:= \frac{1}{|\mathcal{Q}'|}\sum_{q\in\mathcal{Q}'} \mathrm{AUROC}_q. \end{aligned} \tag{23}$$

where i, j are independent draws from the N initial samples on problem q, and the reported r_C, r_W, D are means of $r_{C,q}, r_{W,q}, D_q$ over $\mathcal{Q}'$.

AUROC computation. We compute Eq. (23) as the trapezoidal area under the empirical ROC on each problem's N initial samples, with ties contributing $1/2$. Only problems containing both correct and wrong samples enter $\mathcal{Q}'$. Tables 2 and 1 average AUROC and reproduction rates over this same subset, weighting problems equally rather than by their numbers of cross-class pairs.

For prefix consistency at $K = 1$, the binary score evaluated at the initial answer, $c_i(a_i) \in \{1/2, 1\}$, has only one non-trivial operating point on the false-positive/true-positive rate plane, $(\mathrm{FPR}, \mathrm{TPR}) = (r_{W,q}, r_{C,q})$, so the trapezoidal area gives $\mathrm{AUROC}_q = (1+D_q)/2$, with D_q expressed as a probability difference rather than percentage points. Tables 1 and 2 therefore encode the same information for prefix consistency, and the $\overline{\mathrm{AUROC}}$ ordering across benchmarks matches the D ordering.

G.3 Cost-Accuracy Evaluation (Sections 4.2 and 4.3)

Sample-with-replacement trials. At each token budget B, we draw samples from the pool with replacement and retain the longest prefix of draws whose cumulative cost is at most B. The next draw is excluded if it would exceed B. Initial samples and PC groups are indivisible sampling units. If no unit fits, the method returns no answer and is scored wrong. Exact vote ties are resolved by first insertion into the vote accumulator, hence by answer encounter order in the sampled sequence. We report mean accuracy over M trials. The trial-mean variance used to derive the CI is $\sigma^2 = \frac{1}{M|\mathcal{Q}|^2}\sum_q \hat{p}_q(1-\hat{p}_q)$, where $\hat{p}_q$ is the trial-mean correctness on problem q and $|\mathcal{Q}|$ is the number of problems. Here, σ is the Monte Carlo standard error of the estimated mean accuracy,

and the reported interval is $\pm 2\sigma$. These intervals are conditional on the fixed benchmark questions and generated sample pools: they quantify analysis-side Monte Carlo error, not uncertainty from sampling new questions or generating new pools. This interval is what is shown as "$\pm$" in Table 3.

Dense token-budget grid. Token-efficiency ratios in Table 4 require evaluating the cost–accuracy curve at arbitrary target budgets, so we evaluate every method on the log-uniform grid

$$\mathcal{B} = \left\{10^{3+k/100} \;:\; k = 0, 1, \ldots, 400\right\}.$$

The dense grid is used for all methods that admit a continuous budget (Standard MV, PC variants, DeepConf, CISC, P(True), Response probability). AC and ESC instead contribute their natural-stopping points (one per swept hyperparameter) as described in Appendix E.

Standard MV plateau and Pass@1. The Standard MV plateau is Standard MV's bootstrap-saturated accuracy on the N-sample initial pool, taken as Standard MV's stored accuracy at $\max \mathcal{B} = 10^7$ tokens. We use this finite-budget anchor rather than the unbounded i.i.d. asymptote (which is pool-determined but, on slow-converging problems, can sit above what any finite budget reaches) so that the target stays reachable by Standard MV at every $\alpha \in [0, 1]$. Pass@1 is the closed-form expected accuracy of a single uniformly drawn pool sample, computed directly from the pool.

Monotone envelope and log-budget interpolation. Each method's grid of (budget, accuracy) points is reduced to its running-max envelope along sorted budget (so accuracy is non-decreasing in budget). This gives the "min budget at which the method has ever reached this accuracy" semantics that the token-efficiency ratio is meant to compare. To read off the budget at a target accuracy $\mathrm{acc}_{\mathrm{tgt}}$, we find the first envelope segment $(B^{(j-1)}, \mathrm{acc}^{(j-1)}) \to (B^{(j)}, \mathrm{acc}^{(j)})$ that brackets $\mathrm{acc}_{\mathrm{tgt}}$ and linearly interpolate in (accuracy, log-budget):

$$\log B_{\mathrm{tgt}} = \log B^{(j-1)} + \frac{\mathrm{acc}_{\mathrm{tgt}} - \mathrm{acc}^{(j-1)}}{\mathrm{acc}^{(j)} - \mathrm{acc}^{(j-1)}} \cdot (\log B^{(j)} - \log B^{(j-1)}).$$

The same rule applies to AC and ESC, whose curves are the natural-stopping point lists.

Parametric bootstrap for ratio CIs. Confidence intervals on $B_{\mathrm{method}}/B_{\mathrm{MV}}$ are the point estimate ± 2 standard deviations of the ratio distribution from a parametric bootstrap with the same trial Monte Carlo noise model used by Table 3. Each draw perturbs every fixed-budget accuracy entry by independent $\mathcal{N}(0, \sigma^2_{\mathrm{acc}})$ noise with $\sigma_{\mathrm{acc}} = \mathrm{CI}_{\mathrm{acc}}/2$ (the Monte Carlo standard error from the stored ± 2 standard-error interval), and additionally perturbs each natural-stopping operating point's cost by $\mathcal{N}(0, \sigma^2_{\mathrm{cost}})$ noise with $\sigma_{\mathrm{cost}} = \mathrm{CI}_{\mathrm{cost}}/2$ (the trial-MC standard error of the natural-stopping cost, which varies across operating points unlike the fixed-budget cost). The plateau is perturbed by an analogous draw using its own CI. Each replicate then recomputes the envelope, the target $\mathrm{acc}_{\mathrm{tgt}} = \mathrm{Pass@1} + \alpha \cdot (\mathrm{plateau}' - \mathrm{Pass@1})$ with the perturbed plateau $\mathrm{plateau}'$, and the ratio. Pass@1 is closed-form over the pool (deterministic) and is held fixed across draws. Anchoring the plateau to Standard MV's stored accuracy at $\max \mathcal{B}$ avoids defining that anchor by a maximum over noisy points. It does not remove selection bias in the running-max crossing curves or preserve cross-budget and cross-method dependence. The ratio intervals are therefore approximate conditional Monte Carlo summaries, without validated nominal coverage. Cells where the method's monotone envelope does not reach the target on the point estimate are reported as "N/A". The CI subscript is additionally suppressed (point estimate shown alone) when fewer than 50% of bootstrap draws reach the target.

G.4 Reproduction-Rate GLM (Section 4.4)

Logistic GLM and cluster bootstrap. The slopes $\beta(r_C)$ and $\beta(r_W)$ in Section 4.4 and Table 19 are the Pass@1 coefficients of a per-(model, category) logistic GLM fit on the expanded per-trial Bernoulli outcomes: $\mathrm{logit}\, r_C$ (and separately $\mathrm{logit}\, r_W$) is fit as an affine function of Pass@1, where each (correct initial sample, regeneration) pair contributes one trial to the r_C fit, each (wrong initial sample, regeneration) pair contributes one trial to the r_W fit, and every trial carries the source problem's Pass@1 as its covariate. The fit uses statsmodels' `GLM` with a binomial family. CIs and p-values come from a joint cluster bootstrap over whole problems: each resample draws problems with replacement (preserving each problem's full set of trials) and refits the GLM. We report 2σ-equivalent percentile intervals for coefficients and pointwise percentile bands for fitted curves, and two-sided p-values from the empirical sign distribution.

H Model Output Postprocessing

We extract final answers from `\boxed{}`, normalize them, and then assess answer equivalence. Because AIME 2025 answers are integers, we use exact matching. For HMMT Feb 2026, Brumo 2025, and FrontierScience-Olympiad, the same answer can be expressed in different forms, so we use an LLM judge to assess equivalence. Table 21 in Appendix G.1 summarizes the answer formats and scoring rules. For verbal confidence, we elicit an additional response and parse the model's self-reported score on a 0–100 or binary scale. For P(True), we instead use token probabilities from the binary confidence response.

H.1 Answer and Verbal Confidence Extraction

Boxed-answer parses. Failed boxed responses receive no vote in any aggregator (Standard MV, PC, DeepConf, CISC). Excluding them reduces the usable pool and can change its answer distribution if failures correlate with answer identity or correctness. Across all 20 cells, the pooled extraction-failure rates are 0.25% for initial generations and 0.42% for regenerated continuations. Seventeen of the twenty cells have failure rates below 1% for both initial generations and continuations. The exceptions are the three Nemotron3-30B cells on FrontierScience-Olympiad, AIME 2025, and Brumo 2025. Only Brumo also exceeds 1% for initial generations (Table 22). On cells with multiple truncation fractions, boxed-extraction failure rates vary by at most 0.08 percentage points across $\tau \in \{0.25, 0.50, 0.75\}$. We treat boxed parse failures as missing votes rather than as a separate signal because the rate is small enough that it does not move the relative comparisons reported in the paper.

Table 22: **Answer- and confidence-extraction failure rates (%).** Generations and Continuations report failures to extract either a boxed answer or a usable numeric fallback from initial generations and regenerated continuations, respectively (τ=0.75, K=1). Rates are pooled over problems and samples. Verbal columns report missing confidence values from secondary calls: 0–100 is the CISC score, Binary the 0/1 correctness verdict, and P(True) the log-probability of the binary call's "1" token. Binary and P(True) share the same call.

			Boxed extraction (%)		Verbal queries (%)		
Model	Dataset	N	Generations	Continuations	0–100	Binary	P(True)
GPT-OSS-120B	FrontierScience-Olympiad	128	0.08	0.05	42.99	12.04	0.00
	HMMT Feb 2026	128	0.00	0.02	47.61	25.14	0.00
	AIME 2025	128	0.00	0.00	48.07	21.59	0.00
	Brumo 2025	128	0.00	0.00	44.87	19.92	0.00
GPT-OSS-20B	FrontierScience-Olympiad	128	0.31	0.35	19.38	8.21	0.00
	HMMT Feb 2026	128	0.00	0.05	26.07	24.76	0.00
	AIME 2025	128	0.03	0.03	55.65	30.60	0.00
	Brumo 2025	128	0.00	0.00	51.67	29.19	0.00
Nemotron3-30B	FrontierScience-Olympiad	128	0.17	1.62	48.77	18.92	0.00
	HMMT Feb 2026	128	0.54	0.59	44.06	18.73	0.00
	AIME 2025	128	0.00	1.17	47.53	20.34	0.00
	Brumo 2025	128	4.14	1.25	42.57	18.90	0.00
Nemotron2-9B	FrontierScience-Olympiad	128	0.03	0.38	0.00	0.03	0.00
	HMMT Feb 2026	128	0.00	0.33	0.00	0.24	0.00
	AIME 2025	128	0.00	0.21	0.00	0.00	0.00
	Brumo 2025	128	0.16	0.08	0.16	0.23	0.16
Ministral3-14B	FrontierScience-Olympiad	64	0.28	0.27	0.02	7.41	0.00
	HMMT Feb 2026	64	0.05	0.05	0.28	6.01	0.00
	AIME 2025	64	0.00	0.00	0.21	10.10	0.00
	Brumo 2025	64	0.00	0.05	0.05	6.98	0.00

Verbal score handling. Verbal 0–100 scans the full secondary completion before its first ")" for the first integer and clips it to $[0, 100]$; extraction fails if no integer is found. Verbal binary parses the first "0" or "1" before the first ")" in the final portion when a CoT-to-final delimiter is present, otherwise in the full response; it fails if neither digit is found. P(True) selects the first post-delimiter position whose top-20 candidates contain "0" or "1", falling back to the first generated position. It normalizes

the probabilities of "1" and "0" when both are present, with single-side fallbacks. Pooled across all 20 cells, confidence-extraction failure rates are 27.31% for Verbal 0–100, 12.82% for Verbal binary, and 0.01% for P(True). Verbal failure rates show wide model spread, under 1% for Nemotron2-9B and Ministral3-14B versus 19%–56% for GPT-OSS and Nemotron3-30B on Verbal 0–100 (Table 22). The AUROC table (Table 2) drops missing scores, while the WMV pipeline replaces them by fixed values of 50 for Verbal 0–100 and 0.5 for Verbal binary and P(True) (Appendix E). Retaining these samples preserves their observed answer frequencies and parse-error rate during resampling; dropping them could change the answer distribution if failures correlate with the boxed answer. Table 23 separately compares dropping missing scores with per-problem median imputation. This sensitivity analysis does not reproduce WMV's fixed-default policy. Its gap is at most 0.031 on every cell, and neither sensitivity mode changes the best signal on any cell.

Table 23: **Verbal-confidence $\overline{\text{AUROC}}$ sensitivity to failure handling.** For Verbal 0–100 and Verbal binary, *drop* excludes samples with unparseable scores, as in Table 2. *Imputed* replaces missing scores with the per-problem median of non-missing scores. This sensitivity variant differs from WMV's fixed defaults (Appendix E). Rows labeled "fail" give missing-score percentages among samples with usable extracted answers.

	GPT-OSS-120B				GPT-OSS-20B				Nemotron3-30B				Nemotron2-9B				Ministral3-14B			
Signal / mode	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo	FSci	HMMT	AIME	Brumo
Verbal 0–100																				
drop	.516	.552	.505	.436	.516	.574	.561	.511	.524	.489	.606	.532	.571	.563	.584	.548	.505	.525	.525	.450
imputed	.510	.536	.501	.448	.514	.559	.550	.506	.520	.491	.579	.534	.571	.563	.584	.548	.505	.525	.525	.450
fail (%)	43.0	47.6	48.1	44.9	19.3	26.1	55.6	51.7	48.7	44.1	47.5	42.4	0.0	0.0	0.0	0.0	0.0	0.3	0.2	0.1
Verbal binary																				
drop	.530	.591	.583	.554	.542	.573	.561	.526	.515	.524	.536	.477	.496	.504	.505	.504	.470	.457	.474	.495
imputed	.527	.560	.555	.529	.542	.564	.549	.523	.515	.520	.544	.478	.496	.504	.505	.504	.475	.458	.471	.496
fail (%)	12.0	25.1	21.6	19.9	8.2	24.8	30.6	29.2	18.9	18.7	20.3	18.3	0.0	0.2	0.0	0.1	7.4	6.0	10.1	7.0

Abbreviations: FSci = FrontierScience-Olympiad, HMMT = HMMT Feb 2026, AIME = AIME 2025, Brumo = Brumo 2025.

Verbal-query token costs. In addition to the initial trace length, the evaluation charges the stored minimum-extraction count. For Verbal 0–100, a recorded count runs through the first closing ")" and can be missing even when a score parses if that delimiter is absent. For the shared binary call, it runs through the confidence-token position selected above and may remain nonzero when text parsing fails. Missing counts are treated as zero, so failed Verbal 0–100 extraction incurs no secondary-call overhead in the reported budget. Output after the extraction point is not charged; this accounting favors the verbal baselines rather than measuring their full incurred generation cost.

Verbal 0–100 parser fragility. The dominant failure pattern is a parenthetical or commentary token landing before the digit, for example "`score) 70`", "`analysis) 87 (commentary?) (final) 93`", "`rating)\n\n**70**.`", "`(C) 80%`", and "`value).\n(If you think...)`". A smaller share of failures is completions with no digit at all, such as "`score). That is this answer: **safe**.`". The model that does not fail at all, Nemotron2-9B, has a median successful response length of just 5 tokens, indicating that it emits "`<digit>)`" and stops. The failing models have median successful lengths of 30–200 tokens, meaning even their successful completions wander past the digit and leave many opportunities for parenthetical content to land before any number. These failures illustrate the difficulty of applying a fixed confidence-response template to reasoning models that interleave thinking, commentary, and LaTeX. More robust elicitation and parsing in this setting remain future work.

H.2 LLM-Based Answer Equivalence Judging

Judge prompts and settings. The FrontierScience-Olympiad prompt follows the official grader of the benchmark [Wang et al., 2025, Appendix B]. The prompt for HMMT Feb 2026 and Brumo 2025 follows MathArena [Balunović et al., 2025]. Both prompts require a "YES" or "NO" verdict on whether two answers are equivalent. We use the evaluated model as its own judge, with the same sampling parameters and chat template as for generation.

Using the evaluated model as the judge is a known limitation. A stronger external grader could change absolute accuracies. Appendix D.6 examines whether our relative comparisons are robust to re-scoring the same generation pool with Claude Sonnet 4.6.

The full prompts used are:

FrontierScience-Olympiad.

```
You are grading an attempted answer to a science olympiad problem. You will
be given the attempted answer and the reference answer. Evaluate strictly,
but fairly. The reference answer is either a single number or expression in
latex formatting, a chemical formula, a compound name, or a phrase referring
to a specific name, entity, or method. Mark the attempted answer as correct
if it fully matches the reference answer or is otherwise equivalent (e.g., an
equivalent algebraic expression, a numerical number within 1 decimal place
rounding of the reference answer (e.g., 6.69≈6.7), an equivalent name for a
compound/formula, equivalent when accounting for units, etc.). Mark it as
incorrect if it is not equivalent to the reference answer.

Attempted answer: {answer1}
Reference answer: {answer2}

Answer only ‘‘YES’’ if correct or ‘‘NO’’ if incorrect.
```

HMMT Feb 2026, Brumo 2025.

```
Determine whether two mathematical answers are numerically identical.

Answer 1: {answer1}
Answer 2: {answer2}

Criteria:
  • If they represent exactly the same numerical value, respond ‘‘YES’’
  • If they represent different values or one is not numerical, respond ‘‘NO’’
  • Different formats (fractions, decimals, radicals, exponential notation) are
    acceptable if numerically equivalent

Examples: ‘‘1/2’’ and ‘‘0.5’’ → YES; ‘‘√4’’ and ‘‘2’’ → YES; ‘‘2^3’’ and ‘‘8’’ → YES;
‘‘3.14’’ and ‘‘π’’ → NO; ‘‘x+1’’ and ‘‘1+x’’ → YES; ‘‘x^2’’ and ‘‘2x’’ → NO.

Answer only ‘‘YES’’ or ‘‘NO’’.
```

I Reproducibility Statement

Code and data. The analysis pipeline and reproduction scripts are released at `https://github.com/naoto-iwase/prefix-consistency`, and the answer pool used to reproduce all reported numbers is released at `https://doi.org/10.5281/zenodo.20082164`. Raw model outputs, including sampled text and log-probabilities of the sampled token and the top 20 candidate tokens at each generation step, are also available. See the code repository above for access details. Models and datasets are publicly available (Table 24). Initial generation is stochastic (no vLLM seed), but aggregation and voting use seed 42, so every reported number is bitwise reproducible from the pool.

Inference and sampling. Each model is served through a vLLM OpenAI-compatible endpoint on $4\times$ NVIDIA A100 80GB GPUs, with context window $131{,}072$ and a maximum output length of $100{,}000$ tokens. Sampling uses each model's recommended settings (Table 25).

Table 26 reports per-generation token counts at $\tau{=}0.75$ for all models and at $\tau \in \{0.50, 0.25\}$ for GPT-OSS-20B, used in the τ-sensitivity sweep (Appendix D.5).

Generation prompts. For the three math benchmarks (HMMT Feb 2026, AIME 2025, Brumo 2025) the system prompt is *"You are a helpful assistant specialized in solving mathematical problems."* For FrontierScience-Olympiad it is *"You are an expert scientist solving olympiad-level problems in physics, chemistry, and biology."* Both append *"Please reason step by step, and put your final answer within `\boxed{}`."* to each problem statement. Initial samples and the regenerated continuations from each truncation point share the same prompt.

Table 24: **Models and datasets used in the experiments.** License names follow the original release pages. Please verify before redistribution.

Name	Reference	License
Models		
GPT-OSS-120B [OpenAI, 2025]	`https://huggingface.co/openai/gpt-oss-120b`	Apache 2.0
GPT-OSS-20B [OpenAI, 2025]	`https://huggingface.co/openai/gpt-oss-20b`	Apache 2.0
Nemotron3-30B [NVIDIA, 2025a]	`https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-BF16`	NVIDIA Open Model License
Nemotron2-9B [NVIDIA, 2025b]	`https://huggingface.co/nvidia/NVIDIA-Nemotron-Nano-9B-v2`	NVIDIA Open Model License
Ministral3-14B [Mistral AI, 2026]	`https://huggingface.co/mistralai/Ministral-3-14B-Reasoning-2512`	Apache 2.0
Datasets		
AIME 2025 [Balunović et al., 2025]	`https://huggingface.co/datasets/MathArena/aime_2025`	CC BY-NC-SA 4.0
HMMT Feb 2026 [Balunović et al., 2025]	`https://huggingface.co/datasets/MathArena/hmmt_feb_2026`	CC BY-NC-SA 4.0
Brumo 2025 [Balunović et al., 2025]	`https://huggingface.co/datasets/MathArena/brumo_2025`	CC BY-NC-SA 4.0
FrontierScience-Olympiad [Wang et al., 2025]	`https://huggingface.co/datasets/openai/frontierscience`	Apache 2.0

Table 25: **Sampling settings per model.** Values follow each model's recommended configuration on its release page.

Model	`temperature`	`top_p`	`top_k`	`reasoning_effort`
GPT-OSS-120B	1.0	1.0	40	medium
GPT-OSS-20B	1.0	1.0	40	medium
Nemotron3-30B	1.0	1.0	20	–
Nemotron2-9B	0.6	0.95	20	–
Ministral3-14B	0.7	0.95	20	–

Table 26: **Average output tokens per generation.** Generations is the original full generation. Continuations are completions from the truncation point onward ($K{=}1$). Verbal queries are secondary completions on each init sample: 0–100 elicits the confidence value, Binary serves Verbal binary and P(True). Values are means of stored actual generated-token counts, excluding missing counts. Verbal-query lengths include output after the extraction point and differ from the minimum-extraction costs charged in the budget evaluation (Appendix E). N is the number of answers per problem.

				Continuations			Verbal queries	
Model	Dataset	N	Generations	$\tau{=}0.75$	$\tau{=}0.50$	$\tau{=}0.25$	0–100	Binary
GPT-OSS-120B	FrontierScience-Olympiad	128	3,341	945	–	–	322	196
	HMMT Feb 2026	128	6,925	2,199	–	–	246	182
	AIME 2025	128	5,442	1,668	–	–	262	220
	Brumo 2025	128	4,782	1,493	–	–	246	190
GPT-OSS-20B	FrontierScience-Olympiad	128	7,813	2,631	4,426	5,906	230	210
	HMMT Feb 2026	128	13,707	4,621	7,992	10,941	295	290
	AIME 2025	128	11,159	3,598	6,232	8,564	273	388
	Brumo 2025	128	9,208	3,056	5,323	7,151	277	308
Nemotron3-30B	FrontierScience-Olympiad	128	29,064	8,333	–	–	1,419	103
	HMMT Feb 2026	128	39,816	12,618	–	–	308	149
	AIME 2025	128	28,941	7,313	–	–	281	119
	Brumo 2025	128	22,243	5,859	–	–	1,158	274
Nemotron2-9B	FrontierScience-Olympiad	128	10,124	3,016	–	–	13	3
	HMMT Feb 2026	128	14,216	4,172	–	–	8	4
	AIME 2025	128	11,715	3,498	–	–	10	3
	Brumo 2025	128	9,634	3,125	–	–	10	5
Ministral3-14B	FrontierScience-Olympiad	64	9,671	5,834	–	–	1,940	1,778
	HMMT Feb 2026	64	8,388	4,560	–	–	2,161	1,768
	AIME 2025	64	7,590	3,735	–	–	1,735	1,451
	Brumo 2025	64	7,739	3,994	–	–	1,802	1,277

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and Section 1 accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theorem 1 is proved in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have all the information needed to reproduce the experiments in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The analysis pipeline and reproduction scripts are publicly available at `https://github.com/naoto-iwase/prefix-consistency`, and the generated answer pool is available at `https://doi.org/10.5281/zenodo.20082164`. Reproduction details are provided in Appendix I.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the experimental setting in Section 4 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report Monte Carlo standard errors and bootstrap confidence intervals, with their construction detailed in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include such information in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents a test-time vote-aggregation method on existing public LLMs and benchmarks, with no direct societal impact beyond that of the underlying models.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We don't release new models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Models and benchmarks are cited and listed with URLs and licenses in Table 24.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and answer pool are documented in Appendix I.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.