COVERAGE IS NOT SELECTION: EXACTLY ATTRIBUTING THE GAINS OF TEST-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling methods (self-consistency, best-of-n under a trained verifier, adaptive sampling) are compared by one number: accuracy at budget n. Per question that number is an exact *product*: *coverage* $C_q(n)$, the probability a correct answer exists among n samples, times *selection*, the probability of returning it given it exists; the first depends on the sampler, the second on sampler and rule. We turn it into an audit instrument: a remainder-free identity splitting any accuracy difference into base-rate, headroom and selection channels; an efficiency $\sigma_R(n)$ normalized between a content-blind floor and the oracle ceiling; and unbiased estimators pricing any rule at any budget $n \leq N$ from one pool of N cached samples. On 149,000 generations from six models it does three things accuracy alone cannot. It *attributes*: a token cap and 63 hours at a fixed configuration leave accuracy unresolved at Q=60 yet move the selection channel (cap: σ 0.60 $\rightarrow$ 0.33; 63 hours: -0.36 at n=32, $p = 0.005$, and -0.17, $p = 0.03$, at n=16; one unreplicated pair). It *prices verifiers*: a 7B process reward model is the best selector we measure, yet selectors lose efficiency as the budget grows on all 9 (cell, selector) pairs (6 resolved, uncorrected), and how fast a vote loses it tracks how concentrated its per-sample weights are (Spearman -0.76 with effective sample size, 17 pairs). It *audits adaptive schedules* at matched expected budget: adaptive-consistency gains on the three cells with ample coverage ($+0.016$ to $+0.022$), though which channel pays is unresolved. All code, per-cell results and cached pools behind every number here are released at `https://github.com/xxxxx/xxxxx` (URL anonymized for review; public upon acceptance); code and results also accompany this submission, the pools being too large for it.

1 INTRODUCTION

When a language model answers a question many times, two things improve at two different speeds. The probability that a correct answer *exists* among n samples, the *coverage* $C(n)$ (Brown et al., 2024), climbs monotonically over orders of magnitude of n. Whether the system *returns* that answer is another matter: a selection rule such as plurality voting (Wang et al., 2023) or best-of-n under a learned scorer (Cobbe et al., 2021; Lightman et al., 2024) must recognize the right sample among the wrong ones, and its ability to do so is capped by properties of the answer distribution that more sampling does not repair (Chen et al., 2024; Bay & Yearick, 2026).

The community measures both through one number: task accuracy at budget n, whether a paper proposes a selection rule, tunes a temperature or advocates an adaptive schedule. Where the two *are* separated it is per domain, never as a general accounting: Chen et al. (2021) plot mean-log-probability reranking against the pass@k oracle and Li et al. (2022) report n@k beside pass@k. Per question, accuracy is an exact *product* of a sampler factor and a sampler-and-rule factor, which can respond in opposite directions to the very interventions (temperature, token cap, sample count) papers change *while* comparing selection methods. A reported accuracy difference can therefore be dominated by a channel other than the one the claim is about, and in configurations we exhibit it is: T=0.6 $\rightarrow$ 1.0 at fixed rule and budget moves accuracy $+0.2$ points, a null superposing a 1.4-point base-rate loss, $+1.8$ points of headroom and -0.2 of selection (§7). This paper turns that observation into a measurement theory with three parts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

An exact decomposition and attribution identity (§3). Per-question accuracy factorizes exactly as $A_{R,q}(n) = C_q(n)\, s_{R,q}(n)$. Writing aggregate accuracy as $A = B + \sigma\, H$, with a content-blind floor B, headroom $H(n) = C(n) - B$, and a normalized *selection efficiency* $\sigma_R(n)$, 0 for any rule ignoring sample contents and 1 for an oracle, gives a symmetric, remainder-free identity splitting the accuracy difference between *any* two configurations into base-rate, headroom and selection channels (Theorem 1). The algebra is elementary; only the third channel is evidence about selection.

Optimal offline estimation (§4). Every accuracy, floor and coverage in the decomposition is estimable from one cached pool of N i.i.d. samples per question: all rules, all budgets $n \leq N$, at zero marginal generation cost. A random n-subset of an i.i.d. pool is itself i.i.d., so subset averaging is unbiased for any rule; averaging over *all* subsets is the Rao–Blackwellization and, for the unrestricted i.i.d. model, the UMVUE, with the classical pass@n estimator (Chen et al., 2021) being its oracle case. For count-only rules a dynamic program evaluates that average *exactly*, replacing the 10-draw resampling default, whose noise we measure at 0.4–0.6 of a typical rule gap.

Three audits (§7). We build pools for six models, three commercial APIs (GPT-4.1-nano, Qwen-turbo and GPT-5-nano, a reasoning model) and three open-weight models (Llama-3.2-3B, Phi-4-mini, Mistral-7B) served locally, on competition mathematics (the 367 level-3–5 problems of MATH-500; AIME 2024–25, $Q = 60$). *Sampler drift.* The regime a cell sits in belongs to the sampler, not the model: the token cap alone moves GPT-4.1-nano on AIME from $\sigma_{\mathrm{MV}}(32) = 0.60$ to 0.33 while accuracy moves an unresolved $+4.9$ points, and re-running the same configuration against the same served endpoint 63 hours later moves it $0.95 \to 0.60$ ($p = 0.005$; at n=16, where the pool doubly covers the budget, $0.82 \to 0.65$, $p = 0.03$) with accuracy unresolved; the same test within one session, possible only on nano/MATH (Appendix E.4), moves -0.031 against a null s.d. of 0.055. *Trained verifiers.* A 7B process reward model is the best selector we measure, yet pointwise selectors, it among them, lose efficiency as the budget grows on all nine (cell, selector) pairs we test. Votes are not a uniform family: how fast one loses efficiency tracks how concentrated its weights are, not its label. *Adaptive schedules.* Adaptive-consistency is worth more than its budget saving: at matched expected budget it beats fixed-n voting on the three cells with ample coverage.

Our stance toward prior work is completion, not contradiction: Brown et al. (2024) named coverage and flagged selection as the open problem and Bay & Yearick (2026) measured the gap on released logs, but the measurement layer is missing: an exact factorization with a calibrated scale, estimators with optimality guarantees, and an identity making "where did the gain come from?" computable. We also run a small, temperature-only version of the decoding sweep Bay & Yearick (2026) leave to follow-on work; the accuracy-optimal temperature rises with the budget, and following it gains resolvedly in two families of three.

2 Related work

Repeated sampling and coverage. Brown et al. (2024) showed coverage grows predictably over four orders of magnitude and coined the term; Schaeffer et al. (2025) explained the power-law shape, and Snell et al. (2025) established compute-optimal sampling/model-scale trade-offs. That line studies the ceiling; we treat its $C(n)$ as one factor and measure the second. It also names what Yue et al. (2025) observe when RLVR-trained models beat their base models at k=1 and lose at large k: a base-rate gain paid for with coverage.

Selection rules. Plurality voting over answer clusters is minimum-Bayes-risk decoding under a $0/1$ answer-equality utility (Bertsch et al., 2023), and the oracle-versus-selected gap σ normalizes is the n-best-versus-1-best diagnostic of translation rescoring (Och et al., 2004); what those lack is the per-question factorization, the calibrated scale and the attribution identity. Self-consistency (Wang et al., 2023), outcome and process verifiers (Cobbe et al., 2021; Lightman et al., 2024), LLM judges and weak-verifier ensembles (Saad-Falcon et al., 2025) all target the selection factor, and recent theory characterizes individual rules: Komiyama et al. (2026) derive voting-ensemble asymptotics, Di et al. (2026) prove minimax-optimality of majority-then-reward, Fu et al. (2025) weight and filter the vote by model-internal confidence, and Huang et al. (2026); Aggarwal et al. (2023); Li et al. (2024) stop sampling early per question, for which σ_R must be indexed by the *expected* budget. Verifier-*weighted* voting beats best-of-n under the same verifier in some settings and not others (Li et al., 2023; Wang et al., 2024a); §7.1 finds the ordering budget-dependent on one cell. These works bound *particular* rules; none accounts rule-generically for where realized accuracy comes from.

Limits and diagnoses of test-time scaling. Stroebl et al. (2024) show resampling under an imperfect verifier is capped by its false-positive rate, the verifier-based case of the selection factor, and Wu et al. (2024) map compute-optimal configurations across budgets. Chen et al. (2024) showed vote accuracy can be non-monotone in the number of calls; Bay & Yearick (2026) separated the "identifiability gap" and introduced the modal ceiling (M_∞ below); Liu (2026) and Nowak (2025) develop estimation theory for vote-accuracy curves, Liu et al. (2024) a stability metric, and Song et al. (2025) formalize the generation–verification gap that governs self-improvement. We adopt the identifiability-gap terminology of the concurrent Bay & Yearick (2026) and differ in three ways: our factorization is per-question and exact rather than a difference of aggregate curves; σ is calibrated between a random floor and an oracle ceiling, supporting coarse regime contrasts and fixed-sampler comparisons though not, as Table 2 shows, a ranking of model families inside MATH; and our attribution identity audits configuration pairs, which is what published claims are.

Decoding parameters. Temperature and nucleus mass (Holtzman et al., 2020) matter for multi-sample inference (Du et al., 2025; Wu et al., 2025). Tempering fixes the modal ceiling but not how the optimal temperature moves with budget (Appendix D).

3 THE COVERAGE–SELECTION DECOMPOSITION

3.1 SETUP

Let Q be a distribution over questions, with $Q(\cdot)$ probability under it; where no confusion can arise we also write Q for a cell's question count. A *sampler* θ, a model with its decoding configuration, induces for each q a distribution $D_{q,\theta}$ over generations y. An answer map $a(\cdot)$ extracts a canonical final answer in a countable space (or $\varnothing$ for abstention), and a grader $c_q \in \{0, 1\}$ marks answers correct, with $c_q(\varnothing) = 0$. Write π_q for the induced answer distribution and $p_q = \Pr_{y\sim D_{q,\theta}}[c_q(a(y)) = 1]$ for the per-question success probability. Given budget n the system draws $Y = (y_1, \ldots, y_n)$ i.i.d. from $D_{q,\theta}$ and a *selection rule* R returns one sample; rules may be randomized and, by exchangeability, are taken permutation-invariant without loss of generality. Throughout θ is suppressed when fixed.

Definition 1. *Coverage is $C_q(n) = 1 - (1 - p_q)^n$ and $C(n) = \mathbb{E}_Q[C_q(n)]$; it equals the accuracy of an oracle verifier and is the population value of pass@n. Rule accuracy is $A_R(n) = \Pr$[returned answer correct], with per-question version $A_{R,q}(n)$. The conditional selection rate is $s_{R,q}(n) = \Pr$[returned correct $\mid K \geq 1$] where K counts correct samples among the n; on questions with $p_q = 0$ the conditioning event is null and we set $s_{R,q} := 0$ (any value would do, since $C_q = 0$ there; the choice only moves mass between the two terms of Theorem 1(b)). Write $\bar{s}_R := \mathbb{E}_Q[s_{R,q}]$ over* all *of Q under that convention, the same measure as $C(n) = \mathbb{E}_Q[C_q]$. The conditional rate $\mathbb{E}_Q[s_{R,q} \mid p_q > 0]$ is the interpretable level but does not satisfy (b); we label it where it appears. Neither affects σ_R, which uses aggregates only. The blind floor is $B = \mathbb{E}_Q[p_q]$; the headroom is $H(n) = C(n) - B$; and the* selection efficiency *of R is*

$$\sigma_R(n) \;=\; \frac{A_R(n) - B}{H(n)} \qquad \textit{whenever } H(n) > 0. \tag{1}$$

Lemma 1 (Floor and ceiling)**.** *Any rule whose choice of index is independent of sample contents has $A_{R,q}(n) = p_q$ for every n, hence $A_R \equiv B$; and every rule satisfies $A_{R,q}(n) \leq C_q(n)$, with equality for the oracle. Consequently, whenever $H(n) > 0$, σ is calibrated: content-blind rules score 0, the oracle scores 1, and $\sigma < 0$ diagnoses an anti-informative selection signal.*

Within a sampler the floor is the natural zero point: B is what "sample n times and pick one at random" achieves at every budget, the single-sample accuracy every rule shares, so any excess over B comes from *reading* the samples. Where samplers abstain differently we use a truncation-matched floor $B_{\text{par}} = \mathbb{E}_Q[\Pr(\text{correct} \mid \text{parseable})]$, and write σ^{tm} for the efficiency it induces; unlike B, no content-blind rule attains it, and $C - B_{\text{par}}$ can be negative. Proofs are in Appendix A.

3.2 EXACT FACTORIZATION AND ATTRIBUTION

Theorem 1 (Decomposition and attribution)**.** *(a) Per question, $A_{R,q}(n) = C_q(n)\, s_{R,q}(n)$ exactly. (b) In aggregate, $A_R(n) = C(n)\, \bar{s}_R(n) + \mathrm{Cov}_Q(C_q(n),\, s_{R,q}(n))$, where $\bar{s}_R = \mathbb{E}_Q[s_{R,q}]$ and Cov_Q are both taken over all of Q under the $p_q{=}0$ convention of Definition 1 (restricting either to $\{p_q > 0\}$*

while leaving $C(n)$ unrestricted breaks the identity). (c) Write any configuration $\kappa = (\theta, n, R)$ as $A(\kappa) = B(\theta) + \sigma(\kappa)H(\kappa)$. For any two configurations,

$$A(\kappa') - A(\kappa) \;=\; \underbrace{B' - B}_{\text{base rate}} \;+\; \underbrace{\tfrac{\sigma+\sigma'}{2}\,\big(H' - H\big)}_{\text{headroom}} \;+\; \underbrace{\tfrac{H+H'}{2}\,\big(\sigma' - \sigma\big)}_{\text{selection}}, \tag{2}$$

an identity with zero remainder. The middle channel is named headroom*, not coverage: it is $\bar{\sigma}(\Delta C - \Delta B)$, so a change leaving C fixed while raising the base rate registers a* negative *headroom channel. Table 3 therefore prints ΔC next to, and outside, the three channels. The symmetric split is the two-player Shapley value (Shapley, 1953) over the changes $\{H \to H', \sigma \to \sigma'\}$, the value singled out by efficiency, symmetry, linearity and the null-player axiom, and the two-factor case of the decompositions of Kitagawa (1955) and Shorrocks (2013). The selection channel does not depend on that choice of players: the three-player game over $\{B, C, \sigma\}$, with $A = (1-\sigma)B + \sigma C$, returns the same $\bar{H}\,\Delta\sigma$ and only regroups the rest as $(1-\bar{\sigma})\Delta B + \bar{\sigma}\,\Delta C$.*

Part (a) is elementary; the contribution is the instrument (c) makes of it. Any claim "configuration κ' beats κ by δ" decomposes into three audited numbers, and only the selection channel supports a claim about better *selection*; comparisons that move temperature, cap or budget alongside the rule mix them, by amounts §7 quantifies. When *only* the rule changes, that channel equals ΔA and the split says nothing new (§7); Appendix E.3 asks how often the split buys a *detection* rather than an attribution, and answers: never. The base-rate channel is ΔB renamed (Theorem 1(c) sets it equal to $B' - B$), so the identity adds two quantities, not three, and on all five sampler-side pairs of Table 4 some raw contrast already excludes zero. What it adds is which of the three moved. Besides $s_{R,q} := 0$ at $p_q = 0$, one convention completes the identity: at zero headroom (n=1) we set $\sigma := 0$, exact, but it forces both channels to $\sigma' H'/2$ *identically* for any sampler, rule and dataset, so single-sample baselines carry no attribution information. The covariance in (b) is mostly the convention: unsolvable questions sit at $(C_q, s_{R,q}) = (0, 0)$, which puts it at 0.136–0.166 over all of Q; restricted to solvable questions it is 0.011–0.031, positive at every temperature (nano/MATH, n=16; Appendix E.3).

Three families of rules recur. *Count-only* rules, meaning plurality voting and any vote weighting a sample only by which answer cluster it lands in, see a drawn subset through its cluster counts alone; for plurality this makes the all-subsets average exactly computable (§4). *Score-based* rules (best-of-n under sequence log-probability, pointwise judges, reward models) return the argmax of a per-sample scalar, so their σ depends on how that scalar ranks samples *within* a question, which aggregate calibration does not imply. *Hybrid* rules use scores as vote weights; they aggregate over clusters but are not count-only, since which samples were drawn changes the weights, so they are estimated by Monte Carlo like every score-based rule, not by the dynamic program. $\sigma_R < 0$ means the signal is anti-correlated with correctness under the current sampler.

4 Estimating the decomposition offline

Everything above is estimable from one *pool* $Y_q = (y_1, \ldots, y_N)$ of i.i.d. samples per question, reused for every rule and budget.

Proposition 1 (Pool estimation)**.** *Let $n \le N$. (a) A uniformly random n-subset of an i.i.d. pool is distributed as n i.i.d. draws; hence for any rule R, subset-averaged correctness is unbiased for $A_{R,q}(n)$: one pool prices all rules at all budgets. (b) The all-subsets average $T_R(Y_q) = \binom{N}{n}^{-1} \sum_{|S|=n} \mathbb{E}_R \mathbf{1}[R(Y_S)$ correct$]$ is the Rao–Blackwellization of (a) and, the empirical measure of an i.i.d. sample from an unrestricted discrete family being complete and sufficient (Lehmann & Scheffé, 1950; Bell et al., 1960), the unique UMVUE. The argument is the classical U-statistic one (Halmos, 1946; Hoeffding, 1948); we claim only its application here and the exact evaluation in (c). The classical unbiased pass@n estimator (Chen et al., 2021) is exactly T_{oracle}. (c) If R depends on samples only through answer-cluster counts (*not *probability-, judge- or PRM-weighted voting or answer mass, whose weights read a per-sample score), T_R is an expectation under a multivariate hypergeometric law; for plurality voting it is computable* exactly *by a dynamic program over clusters in $O(mn^3\tau_{\max})$ time, where $\tau_{\max} \le m$ bounds the tied leading clusters (Algorithm 1), abstentions consuming draws without votes and ties broken uniformly. (d) T_R removes resampling noise but not pool finiteness, which the question-level bootstrap behind our intervals already carries; since $\mathrm{Var}(T_R) \le (n/N)\,\mathrm{Var}($one evaluation$)$ (Hoeffding, 1948), we recommend $N \ge 2n_{\max}$.*

Evaluating maj@n by averaging 10 random subsamples, a common default, has Monte-Carlo standard error 0.0053 at n=16 against rule-vs-rule gaps of 0.009–0.014 (Appendix E.3); that is not the dominant uncertainty (question sampling contributes ≈ 0.008), so the case for the DP is that it removes the term more cheaply than 100 draws, not that resampling is unusable. Score-dependent rules use the subset estimator instead; Appendix B gives the draw counts and the error they leave. We validate the suite on synthetic pools with known ground truth: no estimator shows a pooled bias beyond two standard errors, and the DP matches enumeration of every subset to 10^{-16} (Appendices B, C).

5 Vote asymptotics

One proposition converts the decomposition into a prediction; a second, on decoding temperature, is stated and tested in Appendix D.

Proposition 2 (Vote asymptotics and the efficiency limit)**.** *Let m_q^* be the modal voted answer of π_q* over the non-abstention support *(abstentions consume draws without voting, as in Appendix B; on the truncated nano/AIME cell ∅ would otherwise be the raw mode), with margin $\gamma_q > 0$ over the runner-up. (a) $\left| A_{\mathrm{MV},q}(n) - \mathbf{1}[m_q^* \text{ correct}] \right| \le (|\mathcal{A}_q| - 1)\, e^{-n\gamma_q^2/2}$, two-sided, with $\mathcal{A}_q$ the (finite)* support *of π_q, abstention included, and γ_q the margin on the* unnormalized *answer distribution. (b) Hence $A_{\mathrm{MV}}(n) \to M_\infty := Q(m_q^* \text{ correct})$, the modal ceiling of Bay & Yearick (2026), cf. Komiyama et al. (2026), and with $C_\infty = Q(p_q > 0)$,*

$$\sigma_{\mathrm{MV}}(\infty) \;=\; \frac{M_\infty - B}{C_\infty - B}, \tag{3}$$

which is strictly below 1 exactly when solvable questions with wrong modal answers have positive probability. (c) At a fixed sampler with $\sigma_{\mathrm{MV}} > 0$ at the start of an interval on which coverage grows by $\Delta C > \Delta A_{\mathrm{MV}}/\sigma_{\mathrm{MV}}$, $\sigma_{\mathrm{MV}}(n)$ decreases even where accuracy is non-decreasing; accuracy within ε of M_∞ at both ends ($\Delta A_{\mathrm{MV}} \le 2\varepsilon$) does not *suffice.*

Display equation 3 needs no vote simulated, only the solvable fraction, wrong-mode rate and base rate, but at n close to N its agreement is near-definitional, its in-sample fit is carried by cells too small to resolve anything, and held out it does not beat predicting the median of the measured values (Appendix E.2, Figures 2b and 3).

6 Experimental setup

Samplers. Three commercial APIs (GPT-4.1-nano, Qwen-turbo (Qwen Team, 2024) and GPT-5-nano, a *reasoning* model exposing neither temperature nor log-probabilities) and three open-weight models served with vLLM (Kwon et al., 2023) on one RTX 4090: Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Phi-4-mini-instruct (Abouelenin et al., 2025) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). A sampler is a model *plus* its decoding configuration; the token cap is part of the sampler, truncation-induced abstentions counting as ordinary incorrect behaviour. Caps are 1400 tokens on MATH (12,000 for the reasoning model, whose chain is part of the completion); on AIME, 2000 for GPT-4.1-nano and 3500 for the open-weight families. Because that cap truncates 65% of generations (3–8% for the open families) we regenerated the cell at the matched 3500 cap and, so endpoint drift cannot confound it, at the original 2000 cap in the same session; all three pools are reported.

Throughout, *resolved* means the 95% question-level bootstrap interval excludes zero. **Datasets and pools.** MATH (Hendrycks et al., 2021): a reference set of 90 questions from MATH-500 (Lightman et al., 2024), levels 3–5 (15/33/42), swept over temperature for every family, extended at T=0.8 to the full levels-3–5 pool ($Q = 367$) for four of the six samplers. Which set a number comes from is stated with it: Table 1 and §7.1 use the 367 for nano, Qwen, Llama and Phi; the temperature sweeps, Appendix E.3's rule contrasts and Table 2's controls use the reference 90 (Qwen-turbo covers a 61-question *subset*). The within-MATH ordering is not stable across the two sets, so we do not rank families inside MATH by σ (Table 2). AIME 2024–2025: all 60 problems, base rates 0.002 (Mistral) to 0.60 (the reasoning model at its larger cap). Each cell caches $N = 32$ i.i.d. samples per question; at $T = 0.8$, $N = 64$ on all 367 nano MATH questions and on the reference 90 for Llama and Phi (Appendix E.1 lists N per cell). Proposition 1(d) recommends $N \ge 2n_{\max}$, which at n=32 holds

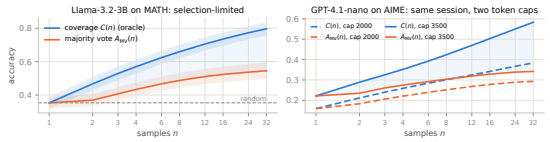


Figure 1: Left: Llama-3.2-3B on MATH (Q=367) is *selection-limited*: coverage (blue) climbs while majority voting (orange) plateaus far below it (gap at n=32: 0.252 $[0.210, 0.291]$). Right: GPT-4.1-nano on AIME at two token caps in one session; raising the cap 2000 $\rightarrow$ 3500 lifts $A_{\mathrm{MV}}(32)$ by an unresolved $+0.049$ while coverage rises $0.383 \rightarrow 0.583$, so $\sigma_{\mathrm{MV}}(32)$ falls $0.60 \rightarrow 0.33$. The accuracy contrast does not exclude zero; the channels do. Bands: 95% question-level bootstrap.

for those questions but not for Qwen, Mistral or GPT-5-nano on MATH, the other 277 Llama and Phi questions, or any AIME cell, where $n = N = 32$ leaves one subset and coverage degenerates to $\mathbf{1}[K \geq 1]$; Appendix E.4 repeats the drift and cap claims at n=16. Nano, Llama and Phi are swept over seven temperatures $\{0.2, \ldots, 1.4\}$, Qwen and Mistral over three; the reasoning family runs at one setting on MATH and two caps on AIME, and it and Mistral stay at the reference 90. Altogether 149,122 distinct generations, \$34.4 of API spend plus \$1.6 of judge scoring (recomputed from cached usage) and about eight RTX-4090-hours; control pools add 116,544 (Appendix E.1).

Grading and clustering. Final answers come from `\boxed{}` spans, graded and clustered by sympy-backed equivalence (Kydlíček & Hugging Face, 2025), so votes operate on semantic answers. We also record sequence log-probabilities where the serving stack exposes them.

Rules. Random pick (the floor), plurality vote (MV), vote weighted by mean token probability, answer probability mass, best-of-n by mean-token-log-probability, LLM-judge scores as argmax and as vote weights, and the oracle, all evaluated offline via §4. The judge is GPT-4.1-nano at T=0 scoring the nano pools, that sampler grading itself, so the two judge rules exist in one family only (prompts in Appendix E.1). Nucleus mass stays at the provider default.

7 Results

Table 1 reports the decomposition at the reference temperature. On MATH the gap $C(32) - A_{\mathrm{MV}}(32)$ excludes zero for every non-reasoning family and is five times larger for the open 3–4B models (0.19–0.25) than for the two non-reasoning APIs (0.03–0.06), with disjoint intervals. Coverage does not explain it (Phi's 0.892 exceeds nano's 0.861), but neither does the mode alone: the gap is $H(1-\sigma)$, so Phi over nano is $3.4\times = 2.2 \times 1.6$ and Llama over Qwen is $7.8\times = 3.9 \times 2.0$, headroom the larger factor in both because the floors differ (0.526 vs. 0.691, 0.353 vs. 0.757) at similar coverage; against the truncation-matched floor the selection factors nearly equalize ($1.05\times$, $1.11\times$) and headroom carries almost all of it (Appendix E.3). AIME reverses the picture, and shows the regime is owned by the sampler, including parts nobody chose. Our first nano/AIME pool, at a 2000-token cap, had a gap of 0.008 $[0.000, 0.025]$ and $\sigma_{\mathrm{MV}}(32) = 0.95$, voting at the oracle ceiling. It does not reproduce. Re-running the identical configuration against the same endpoint 63 hours later (per the pools' timestamps) gives 0.60 $[0.38, 0.84]$ and a gap of 0.090 $[0.028, 0.161]$.

Because the two pools are draws from the same nominal sampler, the null "nothing changed" has an *exact* test: the 64 samples a question holds across them are exchangeable, so every 32+32 split is equally likely. Permuting within question (10^5 draws, both pools clustered jointly so no grading artefact leaks in) gives each difference its null; signs are rerun minus original. What accuracy sees is unresolved ($\Delta A_{\mathrm{MV}} = -0.032$, $p = 0.130$; $\Delta B = +0.002$, $p = 0.834$; $\Delta C(32) = +0.050$, $p = 0.147$); what moves is the efficiency, $\Delta\sigma_{\mathrm{MV}}(32) = -0.356$ against a null s.d. of 0.133 ($p = 0.0045$), with the raw gap $\Delta(C - A) = +0.082$ ($p = 0.0055$), the only two of six statistics clearing a Bonferroni threshold of $0.05/6$. The contrast is budget-indexed, tested as an interaction on the same permutations rather than inferred from two p-values ($p = 0.011$, nominal: it misses a Bonferroni threshold over seven statistics): $\sigma_{\mathrm{MV}}(n)$ climbs $0.72 \rightarrow 0.95$ between n=8 and n=32 on the original pool and drifts from 0.65 to 0.60 on the rerun, the difference growing from an unresolved -0.073 through -0.174 at n=16 ($p = 0.029$, nominal) to -0.356. The one budget-independent difference is the wrong-mode rate, $0.05 \rightarrow 0.22$ (one solvable question of 20 against five of 23, $p = 0.054$), and a wrong mode only entrenches as n grows. Three controls and a caveat

Table 1: The decomposition across six models and two datasets at T=0.8 (T is not exposed by the reasoning family), n=32. $N = 64$ on all 367 nano MATH questions and on the reference-90 questions of Llama and Phi, $N = 32$ elsewhere; where N=32 the coverage estimator degenerates to $\mathbf{1}[K \geq 1]$. Q_s counts questions solved by at least one cached sample; Gap $= C(32) - A_{\mathrm{MV}}(32)$; W-md = wrong-modal-answer share among solvable questions; 95% question-level bootstrap. The $\sigma_{\mathrm{MV}}(32)$ column uses the raw floor and is a *within-sampler* reading; Table 2 gives σ^{tm}, under which the within-MATH ordering changes (Spearman -0.60). Subscripts on the AIME rows are token caps; *two pools from one later session; †fast clustering path; ‡a reasoning model; §with Q_s=3, 5.1% of resamples give $\hat{H} = 0$ and the Fieller interval is unbounded (Appendix E.8). A bracket collapsed to a point (cap-4000 reasoning row) is an in-sample statement: with Q_s=24 a zero wrong-mode count is consistent with a population rate near 0.12. The first two AIME rows are the *same* sampler 63 hours apart; the second and third are the controlled cap contrast (§7).

Sampler	Dataset	Q	Q_s	B	$C(32)$	$A_{\mathrm{MV}}(32)$	Gap [95% CI]	$\sigma_{\mathrm{MV}}(32)$ [95% CI]	W-md
GPT-4.1-nano	MATH	367	322	0.691	0.861	0.805	0.056 [0.039, 0.076]	0.67 [0.58, 0.75]	0.08
Qwen-turbo	MATH	367	320	0.757	0.872	0.840	0.032 [0.016, 0.052]	0.72 [0.59, 0.85]	0.04
Llama-3.2-3B	MATH	367	295	0.353	0.797	0.545	0.252 [0.210, 0.291]	0.43 [0.37, 0.50]	0.32
Phi-4-mini	MATH	367	331	0.526	0.892	0.702	0.190 [0.153, 0.231]	0.48 [0.41, 0.56]	0.22
Mistral-7B†	MATH	90	35	0.074	0.389	0.222	0.167 [0.094, 0.245]	0.47 [0.29, 0.65]	0.43
GPT-5-nano‡	MATH	90	89	0.903	0.989	0.956	0.033 [0.000, 0.078]	0.61 [0.22, 1.00]	0.03
GPT-4.1-nano$_{2000}$	AIME	60	20	0.157	0.333	0.325	0.008 [0.000, 0.025]	0.95 [0.85, 1.00]	0.05
GPT-4.1-nano$^{*}_{2000}$	AIME	60	23	0.159	0.383	0.293	0.090 [0.028, 0.161]	0.60 [0.38, 0.84]	0.22
GPT-4.1-nano$^{*}_{3500}$	AIME	60	35	0.221	0.583	0.342	0.242 [0.142, 0.350]	0.33 [0.17, 0.52]	0.46
Llama-3.2-3B	AIME	60	13	0.028	0.217	0.064	0.153 [0.067, 0.250]	0.19 [−0.02, 0.46]	0.69
Phi-4-mini	AIME	60	14	0.037	0.233	0.062	0.172 [0.083, 0.268]	0.13 [−0.07, 0.38]	0.79
Mistral-7B†	AIME	60	3	0.002	0.050	0.000	0.050 [0.000, 0.117]	−0.03 [unbdd]§	1.00
GPT-5-nano$^{‡}_{12000}$	AIME	60	47	0.604	0.783	0.767	0.017 [0.000, 0.050]	0.91 [0.71, 1.00]	0.02
GPT-5-nano$^{‡}_{4000}$	AIME	60	24	0.302	0.400	0.400	0.000 [0.000, 0.000]	1.00 [1.00, 1.00]	0.00

(Appendix E.4): the two pools truncate 0.654 and 0.647 of generations, so this is not a change in how often the sampler ran out of tokens; the same test at the same budget *within* one session, run on nano/MATH, the only API cell with 64 samples per session, gives -0.031 against a null s.d. of 0.055; between two fresh local sessions of each of Phi-4-mini and Llama-3.2-3B on MATH it gives -0.023 and $+0.037$ against a null s.d. of 0.034, no statistic reaching $p < 0.2$, while changing one server setting (the window vLLM serves Phi with, $4096 \to 8192$ tokens, which switches on its long-context RoPE factors) moves its base rate by -0.098 and plurality accuracy by -0.085 (both $p < 0.001$), so the test detects a sampler that did change, through its base rate; and the cause cannot be named from outside: identical requests to an *unchanged* endpoint already vary run to run, much of it batch-size-dependent kernels under load (Atil et al., 2024; He & Thinking Machines Lab, 2025). The test shows the two sessions differ, not that the difference grew with elapsed time; no check on ΔA would have noticed. That served models change (Chen et al., 2023) and these benchmarks are unstable (Hochlehnert et al., 2025; Miller, 2024) is known; the decomposition adds *which factor* moved.

Against that same-session baseline, raising *only* the cap to 3500 moves accuracy an unresolved $+4.9$ points while coverage rises $0.383 \to 0.583$ ($[0.100, 0.317]$), the wrong-mode rate goes $0.22 \to 0.46$ and $\sigma_{\mathrm{MV}}(32)$ falls to 0.33 $[0.17, 0.52]$ (Figure 1). Truncation is not neutral: it hid hard problems before they could be answered wrongly, invisibly to accuracy, and by point estimate most of the fall in σ is those problems entering headroom rather than voting getting worse on the others (Appendix E.4). Repeating the contrast on GPT-5-nano, a reasoning model, at caps $12{,}000 \to 4000$ in one session takes coverage down 0.383 $[0.267, 0.500]$ and accuracy down 0.367, both resolved, while $\sigma_{\mathrm{MV}}(32)$ rises to *exactly* 1.00: every solvable question then has a correct plurality, because the model answers only what it gets right. A perfect selection efficiency can be manufactured by breaking the sampler, the sharpest form of the floor caveat of §8. Llama's and Phi's AIME wrong-mode rates are 0.69 and 0.79: their errors *correlate* and voting recovers only 0.13–0.19 of the headroom (0.21–0.26 at n=16), as the limit of Proposition 2 predicts in sample.

Attribution audits. The same identity applied to six configuration pairs (Tables 3 and 4) resolves a channel where ΔA does not on four of the five rows where more than the rule changes (though

never where *every* raw contrast fails), and on both nano/AIME contrasts selection is among them (on the cap pair headroom resolves too, with the opposite sign). The sharpest pair: $T{=}0.6 \to 1.0$ at fixed rule and budget moves accuracy $+0.18$ points, the reading "temperature does not matter", by superposing a -1.42-point base-rate channel on a $+1.78$-point headroom channel and a -0.18-point selection channel.

7.1 A TRAINED VERIFIER, ON THE SAME SCALE

The rules so far are verifier-free; most of the literature we audit is not. We scored the MATH pools of four families with Qwen2.5-Math-PRM-7B (Zhang et al., 2025) (run as its card prescribes under `transformers` 4.53.3, reproducing its worked example to within 0.03; AIME scores in Appendix E.5) and added best-of-n under the PRM score and PRM-weighted voting. The verifier ranks samples well (AUC 0.764 nano, 0.720 Qwen, 0.815 Llama, 0.860 Phi, within question) and is the best selector we measure: it beats plurality voting resolvedly at every budget we measure on Llama and Phi, and at $n{=}3$ and $n{=}4$ on nano and Qwen, where above $n{=}8$ the two are indistinguishable (the paired difference at $n{=}32$ is -0.009 $[-0.108, +0.093]$ on nano and -0.010 $[-0.143, +0.123]$ on Qwen, so we do not read a crossing off the point estimates). Verifier rules see the same 32 cached samples per question as the votes, on all $Q = 367$ questions of every cell, and the judge rules appear only on the nano cell, the one family the judge scored.

Pointwise selectors lose efficiency as the budget grows (Figure 5; Table 6 has every value). Of the 9 (cell, selector) trends $\sigma_R(3) - \sigma_R(16)$ in that table, 9 fall as predicted, six resolvedly. We do not pool that with the votes, where no direction is predicted (8 of 17 rise, chance).

Rules within a cell share samples, so rather than count signs we test one paired statistic per cell: the mean selector trend minus the mean vote trend, over *all* rules in each family, exceptions included, bootstrapped over that cell's questions. It excludes zero on two of four cells between $n{=}4$ and $n{=}16$ (two after a Benjamini–Hochberg correction over the 38 trend and contrast tests behind this subsection) and on three of four between the odd budgets $n{=}3$ and $n{=}15$, where a two-way tie needs an abstention (Table 5; on all four fresh open-weight pools and on Llama-3.1-8B it excludes zero at both windows, uncorrected; robustness checks in Appendix E.5). But an odd budget thins out two-way ties only, not three-way ones, and both windows start where plurality ties are common: between $n{=}8$ and $n{=}16$, where ties of any order are rare, the paired contrast excludes zero on no cell ($+0.014$, -0.013, $+0.015$, and $+0.002$), so the resolved claim is one-sided: the selector family's mean efficiency falls on nano, Llama, and Phi at every window we test, $+0.049$, $+0.058$, and $+0.024$ on nano at $(4, 16)$, $(3, 15)$ and $(8, 16)$, likewise on Llama and Phi, while the vote family's mean moves on no cell at $(4, 16)$ or $(3, 15)$ and on one at $(8, 16)$ (Phi, $+0.026$). A family mean is the wrong summary on the vote side, and Llama shows why: all four of its $(3, 16)$ vote trends exclude zero, in both directions, so their mean of -0.014 is resolved movements cancelling, not an absence of movement. What orders the votes is not the vote-versus-selector label but weight concentration: over the 17 (cell, vote) pairs, the effective sample size (ESS) of the weights ranks the $(3, 16)$ trend at Spearman -0.76 ($p = 0.0004$), and -0.73 and -0.58 at $(4, 16)$ and the tie-rare $(8, 16)$. The *sign* is a $(3, 16)$ statement: a cut $\mathrm{ESS} < 0.65n$, read off these pairs, agrees with it on 14 of 17 (of the eight votes that move resolvedly, all but PRM-weighted voting on Qwen), but on 8 at $(8, 16)$, where most votes fall. A weighted vote is a vote only insofar as its weights are spread; as they concentrate it behaves as a selector and decays like one. Confidence-filtered voting (Fu et al., 2025), run on fresh pools, bounds this (Appendix E.5): keeping the top 10% it trends like a selector, keeping 25–90% like plurality, since what it keeps grows with n. On Qwen, whose headroom is the smallest ($H(16) = 0.101$, against 0.15 on nano, 0.38 on Llama, and 0.32 on Phi), neither family contrast resolves, though best-of-n under the PRM ($+0.075$) and PRM-weighted voting ($+0.077$) decay, nominally: with little headroom the instrument is weak. Accuracy rises for every rule but answer mass on Phi.

Within a cell the fall tracks how well the scorer ranks. On nano, where all three scorers exist, the within-question AUCs are 0.764 (PRM), 0.736 (judge) and 0.577 (mean token log-probability) and their decays run in the opposite order: $+0.026$, $+0.044$ and $+0.110$ $[+0.040, +0.177]$, though only the mean-logprob selector and the PRM on Qwen, Llama, and Phi exclude zero alone. The ordering does not extend across cells: the PRM ranks *better* on Llama and Phi (AUC 0.815, 0.860) and decays *more* there ($+0.066$, $+0.060$); only the sign survives across cells. This is what a bounded false-positive rate predicts (Stroebl et al., 2024), and the budget-indexed form of Gao et al.'s (2023)

reward-overoptimization finding: every extra sample gives a pointwise scorer another chance to be fooled, whereas a vote's errors must agree before they cost anything. PRM-weighted voting is the highest-σ rule we measure on nano at every budget ($\sigma = 0.740$, 0.749, 0.739 at n=4, 8, 16), above best-of-n under the same scores, a comparison Li et al. (2023) and Wang et al. (2024a) report both ways, though its margin over the runner-up is not resolved ($+0.016$ $[-0.016, +0.046]$, $+0.005$ $[-0.028, +0.037]$, and $+0.005$ $[-0.035, +0.046]$), so that is an ordering of point estimates. All of this is at a fixed *sample count*: fix total *compute* instead and Singhi et al. (2025) find self-consistency ahead until a generative verifier has had about $8\times$ the budget. The judge rules carry a confound we cannot remove offline: the judge is the sampler grading itself, so its scores correlate with that sampler's own modal answer, which is what a weighted vote rewards; the PRM shares Qwen-turbo's family, and best-of-n under Math-Shepherd, a PRM of another family, decays on all four cells, three resolvedly (Appendix E.5).

7.2 Adaptive sampling, priced at matched budget

Adaptive-consistency (Aggarwal et al., 2023) and early-stopping self-consistency (Li et al., 2024) stop sampling once the answers agree, and report large budget savings at unchanged accuracy. Both are count-only, so a cached pool prices them without new generation, by simulation, since the budget is a stopping time depending on sample order. We simulate each along 200 random orderings of every 32-sample pool (at their authors' defaults) and compare against fixed-budget MV *at the schedule's own expected budget* (Appendix E.7), so "fewer samples" is not the comparison: the question is whether the reallocation is worth anything. The savings reproduce (adaptive-consistency 1.03–$2.41\times$, early stopping 1.01–$3.29\times$, the low end where the model never reaches consensus).

At matched expected budget the reallocation is worth something on the three MATH cells of the nine with ample coverage (all three for adaptive-consistency, two of three for early stopping): adaptive-consistency $+0.022$ $[+0.009, +0.037]$ on nano, $+0.016$ $[+0.006, +0.025]$ on Llama, $+0.018$ $[+0.008, +0.030]$ on Phi; early stopping $+0.004$ to $+0.016$, 2 of 3 excluding zero. Under a rising accuracy curve that is the original papers' own claim read on the matched-budget axis. The decomposition locates the gain, though not decisively: bootstrapped on the same question draws as the totals, on nano headroom resolves and selection does not ($+0.013$ $[+0.001, +0.028]$ against $+0.009$), the saved samples going to questions the schedule has not resolved; on Llama it is the other way round; on Phi neither resolves; and their paired difference excludes zero on no cell (Appendix E.7).

The sign flips where coverage is scarce: by point estimate the headroom channel is negative, for one schedule or both, on 6 of the nine cells, in every case where coverage *at the schedule's own budget* is below 0.55 (a cut read off these cells), but it resolves on one, where early stopping eats the binding resource (Appendix E.7 names them and shows the split is not question count in disguise).

8 Discussion and conclusion

Boundaries, quantified. Appendix E bounds each headline claim under scale and control: equation 3's fit is in-sample (r falls $0.988 \to 0.58$ on the five cells with ≥ 25 solvable questions, and held out, uncalibrated, it beats no constant); the floor is a choice, truncation turning part of a measured efficiency into answer-*existence* filtering; the drift pair clears a multiplicity correction only at $n = N$; and part of §7.1's efficiency contrast is headroom growth.

What to report. Alongside accuracy, report $(B, C(n), \sigma_R(n))$ from the same pools (free once samples are cached) and attribute gains with equation 2, fixing the floor first: raw B within a sampler, B_{par} across samplers abstaining differently. Little headroom cannot adjudicate selection rules, a σ near 1 should prompt a look at abstention, and the cap is worth 20 coverage points here. Re-measure the sampler beside the rule and say which budget was fixed. *Limitations*: checkable answers, all mathematics but a multiple-choice and a free-form check (Appendix E.6); 60 AIME problems; two PRMs, both MATH-trained, and judge rules on nano only; pools assumed stationary within a session; the reference Phi pool not regenerated on a newer serving stack (Appendix E.4); one unreplicated drift pair, on an endpoint whose internal routing we cannot observe.

Reproducibility statement

The estimation suite, the analysis and generation code, and the per-cell results with their bootstrap intervals accompany this submission as supplementary material. The raw sample pools (149,122 generations with per-request token usage) and the PRM and judge scores are *not* included: they run to some 324 MB, past the supplementary limit, and will be released publicly with the camera-ready. Every quantity the paper reports is in the shipped results except the 63-hour interval, which comes from the pools' request timestamps, and `code/consistency_scan.py` checks the paper's numbers against them without the pools; statistics that resample individual samples (the permutation tests, the jackknife) need the pools to be recomputed. Every figure and every generated table regenerates from the shipped results with fixed seeds and no API access; the two hand-set tables (Tables 2 and 3) are checked against those results by the same scan. Estimator unbiasedness is tested by the synthetic ground-truth harness of Appendix C, which runs in minutes on CPU. Regenerating the pools, costed from the cached usage records, takes $\approx$ \$36 in API calls plus about eight hours on one RTX 4090, though API-served cells need not reproduce in distribution (§7). Software versions matter here: the Qwen PRM's model-card code runs under `transformers` 4.53.3, which we pin, and fails under every later release we tried, and a vLLM context-window flag changes Phi-4-mini's sampler (Appendix E.4); the versions behind the open-weight re-runs and the rescoring (vLLM 0.30.0, PyTorch 2.13, CUDA 13.0) are recorded with the code.

Ethics statement

This work analyzes mathematical problem solving and introduces no new model capabilities, data about individuals, or deployment risks. We note a positive externality: the offline pool-reuse methodology reduces the marginal generation cost of selection-rule research to zero after a one-time generation (score-reading rules still need one scoring pass per scorer).

AI use statement

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data (the synthetic answer pools of Appendix C are drawn from specified distributions by the authors' code, and the model generations analysed throughout are the object of measurement, not synthetic data), to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Abdelrahman Abouelenin et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM settings. *arXiv preprint arXiv:2408.04667*, 2024.

Yong Yi Bay and Kathleen A. Yearick. When more sampling hurts: The modal ceiling and correlation ceiling of test-time scaling. *arXiv preprint arXiv:2606.28661*, 2026.

C. B. Bell, David Blackwell, and Leo Breiman. On the completeness of order statistics. *The Annals of Mathematical Statistics*, 31(3):794–797, 1960. doi: 10.1214/aoms/1177705808.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew R. Gormley. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In *Proceedings of the Big Picture Workshop*, pp. 108–122, 2023.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Łukasz Kaiser, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-majority: Minimax-optimal strategy for pass@k inference scaling. In *International Conference on Learning Representations*, 2026.

Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. In *International Conference on Machine Learning*, 2025.

Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

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

Paul R. Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1): 34–43, 1946. doi: 10.1214/aoms/1177731020.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. `https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/`, September 2025. Accessed 2026-08-22.

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

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling (COLM)*, 2025.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. CaTS: Calibrated test-time scaling for efficient LLM reasoning. In *International Conference on Learning Representations*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Evelyn M. Kitagawa. Components of a difference between two rates. *Journal of the American Statistical Association*, 50(272):1168–1194, 1955. doi: 10.1080/01621459.1955.10501299.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. Best-of-∞: Asymptotic performance of test-time LLM ensembling. In *International Conference on Learning Representations*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Hynek Kydlíček and Hugging Face. Math-Verify: A robust mathematical expression evaluation system. `https://github.com/huggingface/math-verify`, 2025.

E. L. Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation. Part I. *Sankhyā: The Indian Journal of Statistics*, 10:305–340, 1950.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, 2023. doi: 10.18653/v1/2023.acl-long.291.

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 *International Conference on Learning Representations*, 2024.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022.

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

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your LLMs capable of stable reasoning? *arXiv preprint arXiv:2412.13147*, 2024.

Yi Liu. Two calls, two moments, and the vote-accuracy curve of repeated LLM inference. *arXiv preprint arXiv:2605.03379*, 2026.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Robert Nowak. Estimating the self-consistency of LLMs. *arXiv preprint arXiv:2509.19489*, 2025.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL*, pp. 161–168, 2004.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen, et al. Shrinking the generation-verification gap with weak verifiers. *arXiv preprint arXiv:2506.18203*, 2025.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? In *International Conference on Machine Learning*, 2025.

Lloyd S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Anthony F. Shorrocks. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *The Journal of Economic Inequality*, 11(1):99–126, 2013. doi: 10.1007/s10888-011-9214-z.

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for LLM reasoning. In *Conference on Language Modeling (COLM)*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, 2025.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling. *arXiv preprint arXiv:2411.17501*, 2024.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024a. doi: 10.18653/v1/2024.acl-long.510.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, 2024b.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

Yuheng Wu, Azalia Mirhoseini, and Thierry Tambe. On the role of temperature sampling in test-time scaling. *arXiv preprint arXiv:2510.02611*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, pp. 57654–57689, 2025.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

A PROOFS

Lemma 1. If the chosen index I is independent of contents, exchangeability gives $y_I \sim D_{q,\theta}$ marginally, so $\Pr[\text{correct}] = p_q$ for every n; average over Q. The ceiling holds because $\{\text{returned correct}\} \subseteq \{K \geq 1\}$, with equality by construction for the oracle. □

Theorem 1. (a) $\Pr[\text{sel. correct}] = \Pr[K \geq 1] \Pr[\text{sel. correct} \mid K \geq 1]$ since the selected answer can be correct only if some sample is. (b) is the definition of covariance. (c) For reals, $\sigma' H' - \sigma H = \frac{\sigma+\sigma'}{2}(H' - H) + \frac{H+H'}{2}(\sigma' - \sigma)$; expand and cancel. Adding $B' - B$ gives the identity. For the Shapley claim: with two players (the H-change and the σ-change) and value function $v(S)$ equal to the accuracy change when exactly the changes in S are applied, the two orderings' marginal contributions average to precisely the two displayed terms, and the Shapley value is the unique value on two-player games satisfying efficiency, symmetry, linearity and the null-player axiom (the egalitarian split satisfies the first three, so the fourth is needed). Averaging over the six orderings of three players instead, the game over $\{B, H, \sigma\}$ on $A = B + \sigma H$ returns equation 2 exactly, and the game over $\{B, C, \sigma\}$ on $A = (1 - \sigma)B + \sigma C$ returns $(1 - \bar{\sigma})\Delta B$, $\bar{\sigma}\,\Delta C$ and the same $\bar{H}\,\Delta\sigma$. □

Proposition 1. (a) Conditionally on the subset's index set, the selected coordinates of an i.i.d. vector are i.i.d. (b) $T_R = \mathbb{E}[\hat{u}_S \mid Y_q$ as multiset], so Rao–Blackwell applies; the empirical measure of an i.i.d. sample from the unrestricted family of discrete distributions is complete and sufficient (on each finitely supported subfamily the counts form a full-rank multinomial family, and completeness on every such subfamily gives it on the union), so Lehmann–Scheffé gives uniqueness. That T_{oracle} equals the pass@n estimator follows by counting the n-subsets that avoid all K correct samples. (c) For count-based rules the correctness of $R(Y_S)$ depends on S only through the drawn cluster-count vector, which follows a multivariate hypergeometric law; Appendix B gives the DP for plurality. (d) Given the pool, T_R is exact, so its only randomness is the pool's; T_R is a U-statistic of degree n on N observations, whose variance is at most (n/N) times that of its kernel (Hoeffding, 1948), which gives the recommendation. The question-level bootstrap resamples per-question estimates and so carries this variance. □

Proposition 2. (a) For a fixed wrong answer a, the vote deficit $\sum_i (\mathbf{1}[a(y_i) = m^*] - \mathbf{1}[a(y_i) = a])$ is a sum of i.i.d. terms with mean $\geq \gamma_q$ and range $[-1, 1]$; Hoeffding bounds the probability it is ≤ 0 by $e^{-n\gamma_q^2/2}$, and a union bound over wrong answers gives the display (uniform tie-breaking only tightens it). (b) Take $n \to \infty$ in (a) and average; the σ limit follows from Definition 1. (c) Write $\sigma(n) = (A_{\text{MV}}(n) - B)/(C(n) - B)$. At a fixed sampler B is fixed, and over an interval on which C increases, σ decreases exactly when $\Delta A_{\text{MV}} < \sigma\,\Delta C$ with σ its starting value; dividing by $\sigma > 0$ gives the stated hypothesis (with $\sigma \leq 0$ and $\Delta A_{\text{MV}} \geq 0$, σ cannot decrease). Convergence to within ε of M_∞ bounds $\Delta A_{\text{MV}} \leq 2\varepsilon$ but does *not* imply the hypothesis. Take 20% right-moded questions with $p_q = 0.65$ and 80% wrong-moded with $p_q = 0.05$, two clusters each, so $B = 0.17$ and $M_\infty = 0.20$. From n=16 to n=32 exact plurality accuracy moves $0.177 \to 0.192$, within $\varepsilon = 0.023$ of M_∞ at both ends ($\Delta A_{\text{MV}} = 0.014 \leq 2\varepsilon$), and coverage rises by 0.197, yet σ *increases* $0.015 \to 0.032$. The ratio test, not the 2ε test, is what (c) asserts; the same mixture has $\sigma = -0.12$ at n=4, which is why (c) requires $\sigma > 0$. □

Proposition 3 (stated in Appendix D; the counterexample of part (c) is given here)**.** (a) With two clusters of masses $p < w$, $\frac{d}{dT}\log(p^{1/T}/w^{1/T}) = T^{-2}\log(w/p) > 0$, so p_T rises toward $1/2$; coverage $1 - (1 - p_T)^n$ increases in p_T and the margin $\gamma_T = 1 - 2p_T$ falls. The right-moded case is symmetric. (b) $x \mapsto x^{1/T}$ is strictly increasing on $[0, 1]$ for every $T > 0$, so tempering preserves the ordering of π_q and in particular its argmax; hence $\mathbf{1}[m_q^*$ correct], M_∞ and C_∞ are independent of T. $\sigma_{\text{MV}}(\infty) = (M_\infty - B)/(C_\infty - B)$ is not, because B moves with T by part (a): in the mixture below it falls from 0.286 at T=0.2 to 0.189 at T=1.4. (c) *A mixture whose optimal temperature falls with budget.* Take two question types in equal proportion, each with two clusters: type A with correct mass 0.52 (right-moded), type B with correct mass 0.18 (wrong-moded). Both premises of (a) hold: under tempering over $T \in \{0.2, \ldots, 1.4\}$, p_B rises $0.0005 \to 0.253$ while p_A falls $0.599 \to 0.514$. Exact binomial majority-vote accuracy with uniform tie-breaking then puts the accuracy-optimal temperature at 1.4 for $n \in \{1, 2, 4\}$ and at 0.2 for $n \in \{8, 16, 32\}$, it moves *down*. The aggregate base rate also rises between the grid's ends here (0.300 at T=0.2 to 0.384 at T=1.4, not monotonically), so even "temperature buys coverage with base-rate currency" is not universal. Sweeping 2000 random two-type mixtures (right-moded mass $p_A \sim U[0.5, 0.9]$ with weight $w \sim U[0.2, 0.8]$, wrong-moded $p_B \sim U[0.01, 0.5]$ with weight $1 - w$; grid $T \in \{0.2, 0.4, \ldots, 1.4\}$;

optimum at n=32 against n=1), the optimum rises with n in 619 cases, falls in 423, and is flat in 958. The argmax direction requires a single-crossing condition in (n, T) that (a) does not supply. □

B EXACT EVALUATION OF VOTE RULES

Plurality voting, the one count-only rule we evaluate, is evaluated exactly by the DP; every rule that reads a per-sample score (best-of-n under any score, and probability-, judge- and PRM-weighted voting) is evaluated by the unbiased subset estimator of Proposition 1(a) with 4000 draws per question in the temperature sweeps and rule comparisons and 2000 in §7.1, for a Monte-Carlo standard error of 2–4 $\times 10^{-4}$, an order of magnitude below the rule contrasts we interpret; the question-level bootstrap resamples these per-question estimates and so carries this noise too.

Let a question's pool have cluster counts $(c_1, \ldots, c_m)$, correct cluster $*$, and an optional abstention cluster $\varnothing$.

Algorithm 1 *Exact T_{MV} by multivariate-hypergeometric DP*

1. Initialise states $S \leftarrow \{(0,0,0) \mapsto 1\}$, where a state (t, M, τ) records the draws allocated so far, the largest vote count among processed wrong clusters, and how many attain it.
2. For each wrong cluster $i \neq *$ with $c_i > 0$: for every state (t, M, τ) with weight w and every $x \in \{0, \ldots, \min(c_i, n-t)\}$, add weight $w\binom{c_i}{x}$ to: $(t{+}x, M, \tau)$ if $i = \varnothing$ or $x < M$; $(t{+}x, M, \tau{+}1)$ if $i \neq \varnothing$, $x = M$ and $M \geq 1$; and $(t{+}x, x, 1)$ otherwise (i.e. $i \neq \varnothing$ and $x \geq M$ with $M = 0$, or $x > M$). The final case is what keeps $x = M = 0$ from falling through and losing mass.
3. Give the correct cluster the remaining $x_* = n - t$ draws with weight $\binom{c_*}{x_*}$, and accumulate $w\binom{c_*}{x_*}$ if $x_* > M$, $\frac{1}{\tau+1} w\binom{c_*}{x_*}$ if $x_* = M \geq 1$ (uniform tie-breaking), else 0.
4. Return the accumulated total divided by $\binom{N}{n}$.

Each of the m stages holds $O(n^2 \tau_{\max})$ states with $O(n)$ transitions each, i.e. $O(mn^3 \tau_{\max})$ work with $\tau_{\max} \leq m$; runtimes are milliseconds at $N \leq 64$. We verified exactness against brute-force enumeration of every subset on 400 random pools with $N \leq 16$, abstentions and tied clusters included (3684 (pool, n) cases, largest deviation 1.1×10^{-16}), and against 10^5-draw Monte Carlo on the synthetic pools of Appendix C (max deviation 1.3×10^{-3}, inside the Monte-Carlo bound of 4.7×10^{-3}).

C ESTIMATOR VALIDATION ON SYNTHETIC GROUND TRUTH

We simulate 40 synthetic questions with known answer distributions (2–8 clusters with Dirichlet masses, every fifth question forced wrong-moded, plus unsolvable and abstention-heavy cases), draw $P = 400$ independent pools of $N = 32$ for each, and compare pool estimates of every rule at $n \in \{1, 2, 4, 8, 16, 32\}$ against fresh-sample Monte-Carlo truth (5×10^4 draws per cell), giving 240 z-scores per rule. Observed exceedance of $|z| > 3$ is 0.0% (random, answer-mass), 0.4% (max-logprob), 0.8% (MV) and 1.3% (oracle), against a null rate of 0.27%; the closed-form coverage estimator has no cell beyond $|z| = 2.03$. The oracle and coverage checks test different objects even though Proposition 1(b) identifies their estimands: the "oracle" row runs the *subsample* path (the generic Monte-Carlo rule evaluator) against fresh-sample Monte-Carlo truth, with a z denominator combining both simulation errors, while the coverage row runs the *closed form* $1 - \binom{N-K}{n}/\binom{N}{n}$ against the analytic $1-(1-p)^n$ under a deliberately conservative $0.5/\sqrt{P}$ denominator. The closed form is exact by construction; what the oracle row measures is the resampling noise the DP removes. The excess is concentrated in near-deterministic cells where the estimator's own standard error is itself estimated from an almost-constant sample and is therefore too small: re-drawing only the cells that exceeded would be regression to the mean, so the direct test is the *pooled signed* bias against

its own standard error: every rule is within two standard errors of zero, the largest being $|z| = 1.82$ (max-logprob). The hybrid, score-weighted votes are not in the harness; they share the subset estimator the oracle row exercises. We report the exceedance rates rather than a pass/fail threshold, since a fixed tolerance well above the null rate would accept a genuinely biased estimator. The harness ships with the code and runs in about three minutes on CPU.

D TEMPERATURE: WHAT TEMPERING CAN AND CANNOT MOVE

Proposition 3 (What temperature can and cannot move)**.** *For the tempered family $\pi_{q,T} \propto \pi_q^{1/T}$ (two clusters): (a) on questions whose modal answer is wrong, raising T strictly increases $p_q(T)$ toward $1/2$, hence coverage at every n, while shrinking the margin $\gamma_q(T)$; on right-moded questions $p_q(T)$ strictly decreases, lowering the base rate. (b) Tempering is rank-preserving on π_q, so the modal answer, and therefore M_∞ and C_∞, are* invariant *to T; the asymptotic efficiency $\sigma_{\mathrm{MV}}(\infty)$ of equation 3 moves only through the base rate. Any temperature effect on vote accuracy is thus a finite-n effect, mediated by margins. (c) These facts do* not *fix the direction in which the accuracy-optimal T moves with n: Appendix A exhibits a two-type mixture whose optimal T moves* down *as n grows, and over* 2000 *random two-type mixtures the optimum rises in* 31%*, falls in* 21%*, is flat in* 48%.

By part (c), budget-dependence of the optimum is an empirical regularity in our families, not a theorem. Two falsifiable predictions remain: (P1) the gap $C(n) - A_{\mathrm{MV}}(n)$ converges to $C_\infty - M_\infty > 0$ wherever wrong-moded solvable questions have positive probability, though not necessarily monotonically (99% right-moded questions at $p_q = 0.8$ and 1% wrong-moded at 0.01 give a gap falling from 0.16 at n=2 to 0.003 at n=24); (P2) equation 3 orders cells by measured $\sigma_{\mathrm{MV}}(n)$ from pool diagnostics alone (held out, Spearman 0.905; Appendix E.2). We therefore treat temperature as measurement; its signature, the argmax T shifting with n, is what a coarse grid or single-budget comparison cannot see.

D.1 TEMPERATURE MOVES THE CHANNELS DIFFERENTLY; THE OPTIMAL SETTING CAN DEPEND ON n

Figure 4 (Appendix E.3) decomposes GPT-4.1-nano's temperature response on MATH: base rate declines beyond T=0.4 (0.641 $\rightarrow$ 0.546), coverage is nearly flat, peaking at T=1.0 (0.833) before degeneration, and σ is comparatively stable. Proposition 3(b) says the asymptotic ceiling cannot move with T *in the answer-level tempered family*; per-token decoding is not that family, so we check the bridge rather than assume it. It holds only approximately: the modal answer agrees with the T=0.8 mode on 81–86% of nano questions and 64–76% of Llama's (excluding the degenerate T=1.4 column), and nano's wrong-mode rate drifts over 0.082–0.123. M_∞ is T*-stable*, not invariant. The observed shift is also less uniform than one line conveys (Figure 2a): nano steps twice (0.4 $\rightarrow$ 0.6 $\rightarrow$ 0.8, at n=3 and n=8), Phi-4-mini once at n=3, skipping a grid point (0.2 $\rightarrow$ 0.6), and Llama-3.2-3B once at n=24. Bootstrapping the argmax itself, $\Pr[\arg\max_T A_{\mathrm{MV}}(32;T) \geq \arg\max_T A_{\mathrm{MV}}(1;T)]$ is 0.84 (nano), 0.93 (Llama-3.2-3B) and 0.997 (Phi-4-mini) with ties included, and 0.59, 0.89 and 0.92 for a strict rise; the accuracy the shift buys at n=32 is +0.007 $[-0.027, +0.043]$, +0.034 $[+0.009, +0.065]$ and +0.111 $[+0.054, +0.175]$, resolved for Llama and Phi but not for nano, and optimistic for all three because the optimal temperatures are chosen on the same questions. A regularity of these samplers on a 7-point grid, not a law, and its *shape* differs by family (Appendix E.3), so a tuned temperature does not transfer.

E ADDITIONAL EXPERIMENTAL DETAIL

E.1 POOLS, PROMPTS AND DATA QUALITY

Samples per question. On MATH the reference-90 sweeps hold $N = 32$ per question at every temperature except T=0.8, where Llama, Phi and nano hold 64; nano's T=0.4 and 1.2 cells hold between 32 and 64 on some questions (mean 36.9 and 37.6), extra same-session draws that the per-question estimators use like any other. Qwen covers 61 of the 90 at three temperatures and Mistral all 90 at three, both with $N = 32$; GPT-5-nano holds 32. At T=0.8 the 367-question pools hold 64 on every nano question, 64 on the reference 90 and 32 on the other 277 for Llama and Phi,

and 32 for Qwen. Every AIME pool holds 32 per problem. The fresh open-weight sessions of Appendices E.4 and E.5 (A and B for Llama and Phi, Phi's session C, and one session of Llama-3.1-8B) hold 32 on each of the 367 questions. These control pools, with the MMLU-Pro and TriviaQA pools of Appendix E.6 (32 on each of 360 questions per task and model), add 116,544 generations (one RTX 4090-class GPU, vLLM 0.30.0) that the 149,122 of §6 do not include. The 90 reference questions are the 61 of a first, level-stratified draw from MATH-500 (its level mix set after a 12-question pilot, whose questions it excludes) plus 29 added to reach 90 (15/33/42 at levels 3/4/5; identifiers ship with the code); the 367-question cells contain every level-3–5 problem of MATH-500, so no selection is involved there. The count of 149,122 is every generation cached for the pools of §6, once, including a few that enter no reported cell (the leftovers of an early nano sweep, a partial AIME run at T=1.2 and a smoke test).

Prompts and decoding. On MATH and AIME every sampler receives one user message, *"Solve this math problem step by step. End your response with the final answer in the form* `\boxed{...}`. *Problem:* {*problem*}*"*; no system prompt or few-shot examples; the last `\boxed{}` span is the answer, with a fallback to a final "answer is" line. Nucleus mass and top-k are left at each serving stack's default; under vLLM that is the model's own generation config, which sets top-$p = 0.9$ for Llama and nothing for Phi, and the fresh Llama sessions of Appendix E.4, served that way, match the reference Llama pool's mean token log-probability (both -0.199), consistent with the same default there. The reference open-weight pools were served with an 8192-token window (the shipped `serve.sh`); under vLLM 0.30.0 such a window makes Phi-4-mini use its long-context RoPE factors at every position, which lowers its base rate by 0.098 against a 4096 window, yet the reference Phi pool, served with 8192 under an earlier installation, matches neither session A (4096) nor C (8192) of Appendix E.4, so the shipped script does not regenerate it exactly. The judge (GPT-4.1-nano, T=0, at most 4 output tokens) receives *"You are grading a candidate solution to a math problem. Problem:* {*problem*} *Candidate solution (may be truncated; final answer at the end): ...* {*last 1500 characters*} *How likely is the candidate's FINAL answer to be correct? Reply with a single integer from 0 (certainly wrong) to 10 (certainly correct). Reply with only the integer."*

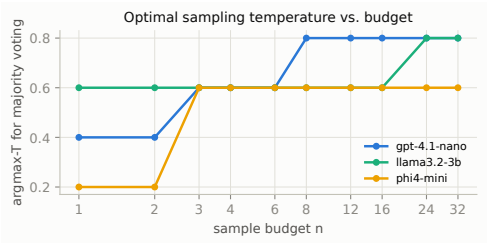
Data quality. Abstention (no parseable boxed answer) and truncation rates per cell and temperature are released with the results. On MATH at the reference temperature abstention is 0.24 (nano), 0.23 (Qwen), 0.10 (Llama), 0.10 (Phi) and 0.05 (Mistral). On AIME, GPT-4.1-nano truncates 0.654 of generations at the 2000-token cap and 0.301 at 3500 (abstention 0.189), against 0.03–0.08 for the open families at 3500. Degeneration at T=1.4 is strongly family-dependent, and it bounds which cells are interpretable: abstention there is 0.28 for nano, barely above its 0.24 baseline, but 0.82 for Llama and 0.88 for Phi, whose T=1.4 columns should be read as degenerate. Table 3 is computed on nano only and its rows stay within $T \in [0.6, 1.0]$.

Grading paths. SymPy-backed equivalence is used everywhere except in one place: the Mistral pools contain many verbose, near-unique answer strings on which pairwise symbolic comparison does not terminate in reasonable time. Those pools alone use a fast path: correctness against the gold answer still uses full symbolic equivalence, while wrong answers are merged by normalised string only. This can only split wrong clusters more finely, which favours the correct answer at the vote; measured on Llama/MATH, where both paths run, the fast path raises $A_{\mathrm{MV}}(32)$ by 0.010 and σ by 0.024. Mistral's σ values in Tables 1 and 2 therefore carry an upward bias of that order, possibly larger, since Mistral's answers are the ones too verbose to compare.

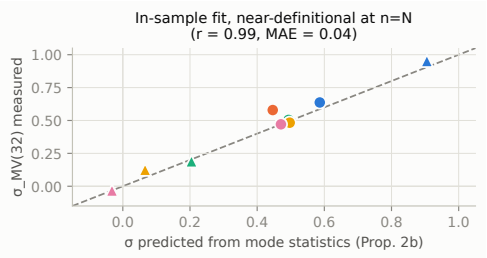
Costs. Every dollar figure in this paper is recomputed from the per-request token usage cached in the pools, each request priced once, by `code/cost_estimate.py` rather than read off a billing counter, so §6, the reproducibility statement and this paragraph quote the same script. API generation: \$34.35 over 8.6M input and 90.2M output tokens (the largest items are the reasoning-model AIME pair and the nano/MATH sweep); judge scoring: \$1.61; total \$35.96.

E.2 The closed form and the landscape

Out-of-sample prediction. Split-half test, all eight cells with a solvable question in each half: diagnostics on samples 1–16, $\sigma_{\mathrm{MV}}(8)$ measured on samples 17–32. Predicted/measured: 0.718/0.532 (nano/MATH), 0.771/0.616 (Qwen), 0.427/0.410 (Llama/MATH), 0.398/0.416 (Phi/MATH), 0.427/0.303 (Mistral/MATH), 1.000/0.649 (nano/AIME), 0.375/0.208 (Llama/AIME), 0.243/0.275 (Phi/AIME). Uncalibrated MAE 0.131 (Pearson $r = 0.910$, Spear-



(a) Argmax temperature for plurality voting against budget, three densely-swept families (Appendix D). Phi's step skips a grid point ($0.2 \to 0.6$), and of the three n=1-versus-n=32 contrasts only nano's is unresolved.



(b) The closed form equation 3 from pool mode statistics vs. measured $\sigma_{\mathrm{MV}}(32)$, *in sample*: near-definitional where $n = N$, shown for completeness; the held-out test is below.

Figure 2: (a) How the voting-optimal temperature moves with budget (Appendix D). (b) The in-sample fit of equation 3 (§5).

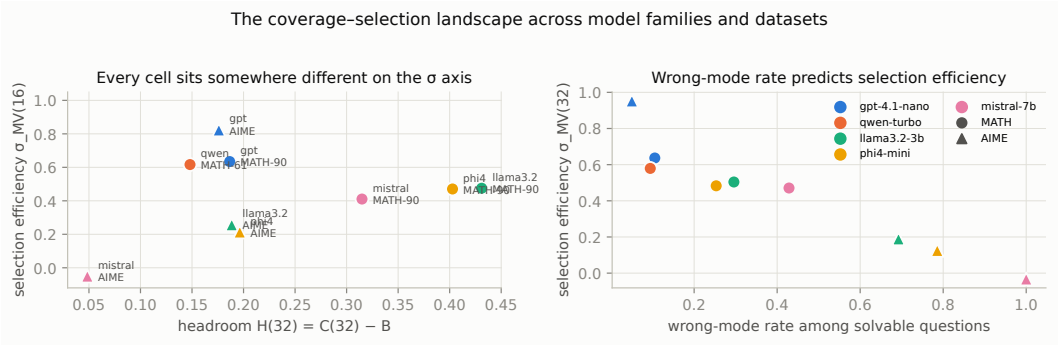


Figure 3: The coverage–selection landscape across nine (sampler, dataset) cells. Left: cells spread over most of the efficiency range at headroom between 0.05 and 0.45 ($\sigma_{\mathrm{MV}}(16)$ against $H(32)$; the right panel is at n=32 throughout; nano/AIME is the original 2000-cap pool). Right: the wrong-mode rate organizes the landscape, as Proposition 2 predicts, though at $n = N$ that relation is close to definitional, since $C(32) = Q_{\mathrm{s}}/Q$ wherever N=32 and a plurality at large n returns the mode; the out-of-sample test is the split-half one below, which equation 3 fails.

man 0.905), against 0.130 for the leave-one-out median of the measured values and 0.148 for their leave-one-out mean; leave-one-out affine recalibration of the prediction 0.072; the same recalibration applied to the wrong-mode rate alone 0.049 (Spearman -0.976). A smaller version of this test, on the three N=64 pools with $\sigma_{\mathrm{MV}}(16)$ on the held-out half, gave predicted/measured $0.702/0.628$, $0.510/0.401$, $0.414/0.470$, MAE 0.080; we quote the larger test.

The closed form, in full. In sample, across the nine cells that predate the cap-matched rerun, $r = 0.988$ and MAE 0.037; *within* MATH alone r falls to 0.58. That 0.988 is a range effect, and dropping cells at the top does not show it: substituting the cap-matched cell gives $r = 0.976$ and dropping both nano/AIME cells gives $r = 0.978$ over the remaining eight, but the leverage is at the *bottom*, where three AIME cells have 3, 13 and 14 solvable questions and a σ whose interval reaches the content-blind floor. Dropping every cell with fewer than 25 solvable questions leaves the five MATH cells, and there $r = 0.58$, Spearman 0.30, and the MAE of 0.042 barely beats the 0.052 of predicting the median. The in-sample fit is carried by the gap between cells that resolve nothing and cells that do. Held out (diagnostics on 16 samples, $\sigma_{\mathrm{MV}}(8)$ on the disjoint 16, eight estimable cells): MAE 0.131, and the paired bootstrap difference against the in-sample median, a constant that favours itself, is $+0.004$ $[-0.031, +0.096]$: uncalibrated, the closed form does not beat predicting the median. It overshoots on 6 of the 8 cells, as a finite-budget target predicts, since $\sigma_{\mathrm{MV}}(8)$ sits below its limit wherever margins are small, by $+0.167$ on average where it does and $+0.119$ averaged over all eight; a leave-one-out affine recalibration brings MAE to 0.072. The wrong-mode rate alone, recalibrated identically, gives MAE 0.049 and Spearman -0.976, beating

Table 2: Truncation-matched floor (§3) and question-set controls, all ten cells at T=0.8, n=32, *on the reference* 90*-question MATH set* (Qwen-turbo on its 61-question subset of it, which is why the last column exists) and on the 60 AIME problems, so the MATH values differ from Table 1's Q=367 rows. B_{par} is the mean over questions of the correct share among *parseable* samples (0 where none parses); σ^{tm} uses it in place of B. "σ on common 61" restricts the MATH cells to the question set Qwen shares (verified to be a subset of the 90). AIME is essentially unaffected and keeps its ordering; MATH does not (Spearman -0.60 raw vs. matched, 0.60 raw vs. common-61).

Sampler	Dataset	B	B_{par}	$\sigma_{\text{MV}}(32)$	$\sigma^{\text{tm}}_{\text{MV}}(32)$	Δ	σ on common 61
GPT-4.1-nano	MATH	0.630	0.693	0.637	0.448	-0.189	0.564
Qwen-turbo	MATH	0.721	0.774	0.579	0.342	-0.238	0.579
Llama-3.2-3B	MATH	0.328	0.347	0.504	0.481	-0.023	0.437
Phi-4-mini	MATH	0.436	0.454	0.483	0.459	-0.024	0.499
Mistral-7B	MATH	0.074	0.075	0.471	0.468	-0.002	0.450
GPT-4.1-nano$_{2000}$	AIME	0.157	0.207	0.953	0.934	-0.019	—
GPT-4.1-nano$_{3500}$	AIME	0.221	0.240	0.332	0.296	-0.036	—
Llama-3.2-3B	AIME	0.028	0.030	0.190	0.181	-0.009	—
Phi-4-mini	AIME	0.037	0.037	0.126	0.124	-0.002	—
Mistral-7B	AIME	0.002	0.002	-0.032	-0.032	$+0.000$	—

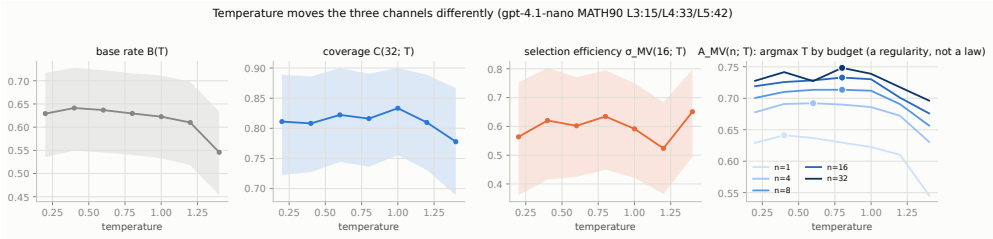


Figure 4: Temperature response decomposed into channels (GPT-4.1-nano, MATH, seven temperatures). Base rate $B(T)$ falls past T=0.4; coverage $C(32;T)$ is nearly flat, peaking at T=1.0; $\sigma_{\text{MV}}(16;T)$ is noisy-stable. Net effect (right): the accuracy-maximizing temperature for voting moves right, from T=0.4 at n=1 to T=0.8 at $n\geq$8, for this family. Across the three densely-swept families the accuracy the shift buys resolves for two (Appendix D).

the closed form in 82% of cell resamples though the contrast is itself unresolved at eight cells (ΔMAE $[-0.085, +0.016]$). Held-out Spearman for the closed form is 0.905, 95% CI $[0.42, 1.00]$, positive in 99.7% of resamples; dropping the truncated cell moves the pair to MAE 0.100 and Spearman 0.857. Within MATH the σ spread is 0.17 raw and 0.14 against the truncation-matched floor, in both cases no larger than the movement either control induces.

E.3 Attribution audits, and what the identity adds

The covariance of Theorem 1(b). On the nano/MATH reference sweep at n=16, $\text{Cov}_Q(C_q, s_{R,q})$ for plurality voting is 0.136–0.166 over the seven temperatures when taken over all of Q, and 0.011–0.031 over solvable questions, every one of the latter intervals excluding zero. The first number is mostly the $s_{R,q} := 0$ convention: with u the unsolvable share, $\text{Cov}_Q = (1-u)\text{Cov}_{\text{solv}} + u(1-u)\,\mathbb{E}_{\text{solv}}[C_q]\,\mathbb{E}_{\text{solv}}[s_{R,q}]$, and the second term alone is 0.11–0.15 here. Its sign carries no diagnostic content by itself: C_q and $s_{R,q}$ both rise with p_q under plurality, so the solvable-question covariance can be positive even when wrong-moded questions dominate.

Does the decomposition ever resolve what the raw contrasts do not? σ and the channels are functions of (B, C, A), so the instrument earns its keep only where a channel's interval excludes zero and the contrasts a reader already has do not. Bootstrapping ΔA, ΔB, ΔC, the gap contrast $\Delta(C-A)$ and the three channels together on the five non-degenerate pairs of Table 3: on all five at least one raw contrast excludes zero, so the answer is *never*. The closest case is the temperature-only pair (intervals from this joint resampling, so they differ from Table 4 in the third decimal): $\Delta A = +0.002$ $[-0.023, +0.025]$, $\Delta C = +0.016$ $[-0.006, +0.041]$ and $\Delta(C-A) = +0.014$

Table 3: Exact three-channel attribution (Theorem 1c) for six configuration pairs (GPT-4.1-nano; rows 1–2 on AIME, the rest on MATH). Rows 1–2 are the §7 contrasts: an intended cap change and an unintended endpoint change, both with accuracy differences that do not exclude zero and selection channels that do. The base-rate, headroom and selection columns sum to ΔA identically (residual 7×10^{-18}; displayed values are rounded, so a printed row may be off by one in the last digit); intervals in Table 4. ΔC sits *outside* the sum: the middle channel is $\bar{\sigma}(\Delta C - \Delta B)$, not ΔC. Bold: selection channels resolved in Table 4; the AIME rows (Q=60) carry one decimal fewer.

Comparison	ΔA	Base rate	Headroom	Selection	ΔC
Cap: 2000 → 3500, one session	$+0.049$	$+0.062$	$+0.064$	$\mathbf{-0.077}$	$+0.200$
Nothing: 2000 → 2000, 63 h apart	-0.032	$+0.002$	$+0.037$	$\mathbf{-0.071}$	$+0.050$
Scale-up: $(0.6, \mathrm{MV}, 8) \to (1.0, \mathrm{MV}, 32)$	$+0.0254$	-0.0142	$+0.0501$	-0.0104	$+0.0717$
Temperature only, n=16: 0.6 → 1.0	$+0.0018$	-0.0142	$+0.0178$	-0.0018	$+0.0156$
Budget only: n=8 → 32 (0.8, MV)	$+0.0348$	0.0000	$+0.0307$	$+0.0041$	$+0.0492$
Rule only: MV → judge-wtd. vote (0.8, 16)	$+0.0062$	0.0000	0.0000	$+0.0062$	0.0000

Table 4: Attribution channels with question-level bootstrap intervals (* excludes zero), for five of the six pairs of Table 3 and a degenerate baseline. ΔB and ΔC are shown beside the channels because the headroom channel is $\bar{\sigma}(\Delta C - \Delta B)$: where it resolves, this says whether it is doing more than restating a resolved coverage or base-rate move. The rule-only row is omitted, because with B and C fixed its selection channel equals ΔA resample by resample; the degenerate row is the n=1 baseline, which forces the two channels equal.

Comparison	ΔA	ΔB	ΔC	Headroom	Selection
Cap, one session	$+0.049\ [-0.033, 0.128]$	$+0.062\ [0.038, 0.092]^*$	$+0.200\ [0.100, 0.317]^*$	$+0.064\ [0.013, 0.124]^*$	$-0.077\ [-0.154, -0.008]^*$
Nothing, 63 h apart	$-0.032\ [-0.078, 0.004]$	$+0.002\ [-0.007, 0.012]$	$+0.050\ [0.000, 0.117]$	$+0.037\ [0.000, 0.077]$	$-0.071\ [-0.122, -0.024]^*$
Scale-up	$+0.025\ [-0.009, 0.063]$	$-0.014\ [-0.027, -0.003]^*$	$+0.072\ [0.030, 0.125]^*$	$+0.050\ [0.027, 0.077]^*$	$-0.010\ [-0.042, 0.023]$
Temperature only	$+0.002\ [-0.024, 0.026]$	$-0.014\ [-0.027, -0.003]^*$	$+0.016\ [-0.006, 0.041]$	$+0.018\ [0.005, 0.032]^*$	$-0.002\ [-0.026, 0.019]$
Budget only	$+0.035\ [0.013, 0.059]^*$	$+0.000\ [0.000, 0.000]$	$+0.049\ [0.025, 0.076]^*$	$+0.031\ [0.015, 0.048]^*$	$+0.004\ [-0.010, 0.020]$
Degenerate: random@n=1 → MV@n=16	$+0.103\ [0.064, 0.145]^*$	$+0.000\ [0.000, 0.000]$	$+0.163\ [0.114, 0.216]^*$	$+0.052\ [0.032, 0.072]^*$	$+0.052\ [0.032, 0.072]^*$

$[-0.013, +0.044]$ all fail while the headroom channel resolves ($+0.018\ [+0.004, +0.033]$), but $\Delta B = -0.014\ [-0.027, -0.003]$ resolves too, and the base-rate channel *is* ΔB. What the channels buy is attribution: on the 63-hour pair $\Delta(C - A) = +0.082\ [+0.025, +0.150]$ says something moved and the selection channel says it was selection; on the temperature pair the one resolved raw contrast says the floor fell, and the headroom channel, together with $\Delta(C - B)$, says the room left to scale into grew. A reader who wants a detection instrument should watch $\Delta(C - A)$, which is free; a reader who wants to know which of three things moved needs the split.

Rule comparisons at a fixed sampler. Fixing the sampler and the budget fixes coverage, so rule comparisons are pure-σ comparisons and, as §3 notes, a positive rescaling of the corresponding accuracy contrasts. Unless marked, every number in this paragraph is (nano, MATH, Q=90, N=64); §7.1's nano values are (nano, MATH, Q=367, N=32) and differ for that reason. Here $\sigma_R(16)$ orders as judge-weighted vote $0.672 >$ MV $0.634 >$ probability-weighted vote $0.615 >$ judge-argmax $0.554 \approx$ answer-mass $0.547 >$ max-mean-logprob 0.442. Paired bootstrap resolves the extremes, not the middle: MV beats the weakest rule by $+0.192\ [0.091, 0.304]$ ($+0.238$ on Phi-4-mini), but against the *best* pointwise selector the margin is less than half as large ($+0.080$) and unresolved at this Q, so we claim that ordering, not its significance. One contrast is robust: the same judge scores are worth more as vote weights than as a selector. It is no taxonomy-level law: answer-mass, which aggregates over clusters just as a vote does, sits at 0.547, and weighted-vote-over-MV flips sign between nano and Phi (below). All six rules here are free of a *trained* verifier; §7.1 adds a trained 7B PRM on the same axis, which is where Stroebl et al. (2024)'s cap enters σ.

Across families. The temperature trade differs in shape across families: for Qwen-turbo the base rate is flat in T ($0.721/0.721/0.722$) while coverage climbs ($0.820 \to 0.885$), so temperature buys coverage almost for free; for GPT-4.1-nano the same move (T=0.4 → 1.2) costs 0.031 of base rate, and 0.095 by T=1.4. Judge scores exist for the nano pools only; Phi-4-mini is scored on the four rules that need no extra calls, and four of its pairwise contrasts exclude zero: MV over max-mean-logprob $+0.238$ (nano $+0.192$), MV over answer-mass $+0.445$, probability-weighted vote over MV $+0.031\ [0.015, 0.048]$, and probability-weighted vote over max-mean-logprob $+0.269$.

The weighted-vote-over-MV ordering is the one that flips: $+0.031$ on Phi against -0.019 on nano, so "weighting helps" is family-dependent even within a single weighting scheme.

Caveats at scale. The gap excludes zero at all seven temperatures for each densely-swept family, but not universally: Qwen at T=0.4 gives 0.046 with an interval touching zero. Several directionally stable rule contrasts are unresolved at Q=90, where the 80%-power minimum detectable effect on σ is ≈ 0.15. Six of the nine landscape cells of Figure 3 have $n = N = 32$, violating our own $N \geq 2n_{\max}$ advice; recomputing at n=16 moves individual values, most notably nano/AIME from 0.95 to 0.82 (σ_{MV}: Qwen 0.62, Mistral/MATH 0.41, nano/AIME 0.82, Llama/AIME 0.26, Phi/AIME 0.21, Mistral/AIME -0.05).

The rule-gap scale. The Monte-Carlo standard error of §4 is quoted against the lower quartile and the median of $|A_R(16) - A_{R'}(16)|$ over the fifteen pairs of the six content-reading rules on the reference nano/MATH cell, 0.009 and 0.014 (the upper quartile is 0.020). It is computed from that cell's results rather than asserted, so the figure in the text and the figure in the results cannot drift apart.

The MATH gap under the matched floor. On the 367-question cells B_{par} is 0.750 (nano), 0.812 (Qwen), 0.372 (Llama) and 0.534 (Phi). Against it the gap factorization of §7 becomes Phi over nano $3.4\times = 3.2 \times 1.05$ and Llama over Qwen $7.8\times = 7.1 \times 1.11$: the selection factor nearly equalizes and headroom carries almost all of the difference, so the raw-floor reading understates rather than overstates the headroom share.

E.4 The drift and cap contrasts

How the drift contrast should be read. The commercial models were accessed through their APIs. A served endpoint is a pipeline (gateway, routing, backend replicas, model revision), no layer of which is observable from the client, so the 63-hour contrast establishes that *the endpoint we called* behaved differently on two occasions; it cannot name which layer moved and is not evidence about a particular vendor's serving stack. The permutation test treats samples as exchangeable within each session, so what it establishes is that the two sessions differ; a session-level effect of any lag would produce the same result, and only a third session separates the two readings. The measurement lesson survives both ambiguities: a configuration repeated against the same endpoint can change regime while ΔA shows nothing. The open-weight families, served locally, admit the check the API cells do not: re-running a sampler known not to have changed (next paragraph).

A fixed sampler, re-run: a local negative control. Phi-4-mini and Llama-3.2-3B were served in two further sessions each with vLLM 0.30.0 on an RTX 4090-class GPU (A and B: separate server processes with different seeds, on one day, identical settings and a 4096-token window), 32 samples on each of the 367 MATH questions with the reference prompt, temperature and cap. These are the *fresh pools* of Appendix E.5. Between A and B (differences B minus A) the within-question permutation test of §7 (5000 permutations, pools clustered jointly) finds nothing: $\Delta\sigma_{\mathrm{MV}}(32) = -0.023$ on Phi and $+0.037$ on Llama, against null s.d. of 0.034 and 0.034 ($p = 0.50$, $p = 0.28$), and $\Delta\sigma_{\mathrm{MV}}(16) = -0.008$ and $+0.022$; over the six statistics of §7 (ΔA_{MV}, ΔB, ΔC, $\Delta\sigma_{\mathrm{MV}}$, $\Delta(C - A_{\mathrm{MV}})$ and the wrong-mode rate), both budgets and both families, the smallest p is 0.20. This is a check on the test where the sampler is known to be fixed, on 367 questions: it does not show that sessions of an API endpoint are exchangeable, which only a third API session could. Against the reference pools (differences session A minus reference, the reference's first 32 samples by index on each question), which were served five to six weeks earlier on another RTX 4090 under an earlier software installation, Llama shows nothing ($\Delta\sigma_{\mathrm{MV}}(32) = -0.047$, $p = 0.15$; smallest p over the statistics 0.11); Phi's base rate moved: $\Delta B = +0.026$ ($p < 0.001$), with $\Delta A_{\mathrm{MV}} = +0.029$ at n=32 ($p = 0.011$) and $+0.029$ at n=16 ($p < 0.001$), while $\Delta\sigma_{\mathrm{MV}}(32) = +0.046$ is unresolved ($p = 0.19$): of the three factors (B, C, σ), only the base rate resolves. The fresh Phi sessions truncate 2.0% and 2.1% of generations against the reference's 4.0%. The reference pools were served by the shipped `serve.sh` with an 8192-token window, and with a window above 4096 tokens vLLM 0.30.0 applies Phi-4-mini's long-context RoPE factors at every position.

A serving setting, as a positive control. A third Phi session, C, identical to A except for that window (8192), isolates it (differences C minus A): the base rate falls by 0.098 ($p < 0.001$) and plurality accuracy by 0.085 at n=32 ($p < 0.001$), while coverage falls by less, 0.034 at n=16 ($p < 0.001$; 0.014 at n=32, $p = 0.31$), so the headroom widens, and the efficiency falls, nominally

at n=32 ($\Delta\sigma_{\mathrm{MV}}(32) = -0.068$, $p = 0.048$) and without resolving at n=16 (-0.038, $p = 0.088$); the wrong-mode rate rises by 0.071 ($p < 0.001$) and truncation goes from 2.0% to 9.7%. A server setting that no model card lists among decoding parameters is thus part of the sampler: it moves Phi's plurality accuracy by 8.5 points, through the base rate, which falls by more than accuracy does (the wider headroom makes up part of it), and the permutation test that is null between sessions A and B detects it in B and A_{MV}, though only nominally in σ. It does not reproduce the reference: session C sits below it ($\Delta B = -0.072$, $p < 0.001$, C minus reference) as A sits above it, so the window alone does not explain the reference-to-fresh difference on Phi, whose cause (the earlier software installation or the other GPU) we cannot identify.

Controls at a non-degenerate budget. At $n = N = 32$ the unbiased coverage estimator reduces to $\mathbf{1}[K \geq 1]$, so $C(32)$ carries no information beyond "some cached sample was correct" and equals Q_{s}/Q exactly in every N=32 row of Table 1. The drift and cap claims are therefore repeated at n=16, where $N \geq 2n$. For the three nano/AIME pools $\sigma_{\mathrm{MV}}(16)$ is 0.824 $[0.720, 0.923]$ (original 2000 cap), 0.649 $[0.477, 0.811]$ (same configuration, 63 hours later) and 0.380 $[0.243, 0.525]$ (3500 cap, same session as the rerun): the drift contrast shrinks from -0.356 to -0.174 (re-running the exact permutation test at that budget gives $p = 0.029$ against 0.0045 at n=32, so it clears 0.05 but not the Bonferroni threshold of $0.05/6$), and the cap contrast is essentially unchanged; both keep their sign and ordering. *Truncation is not the explanation for the drift.* The two 2000-cap pools truncate 0.654 and 0.647 of generations and fail to produce a boxed answer on 0.488 and 0.490, indistinguishable, while the 3500-cap pool truncates 0.301. Against the truncation-matched floor, $\sigma^{\mathrm{tm}}_{\mathrm{MV}}(32)$ is 0.934, 0.488 and 0.296 for the three pools, preserving both contrasts. *The divergence with budget, tested directly.* "Significant at n=32, not at n=8" is not evidence that the contrast grew, so we test the interaction on the same 10^5 permutations: $\Delta\sigma_{\mathrm{MV}}(32) - \Delta\sigma_{\mathrm{MV}}(8) = -0.283$ against a null s.d. of 0.116, $p = 0.011$, clearing 0.05 though not a Bonferroni threshold over the seven statistics that run now carries. *Nor is it within-session variation.* Proposition 1 treats a pool as i.i.d.; if a served endpoint can move over 63 hours it might move over the minutes to hours a pool takes to collect (28 minutes to 12 hours per question here), which would undercut every interval in the paper. Ordering each question's cached samples by request timestamp and splitting them into an earlier and a later half runs the identical within-question permutation test inside one session. The comparison that matters is the one at the *same* budget and the *same* split as the between-session test, since the effect is claimed to grow with budget. Any pool holding $N = 64$ samples from a single session admits exactly that design (earliest 32 against latest 32, evaluated at n=32, coverage degenerate in both arms just as it is between sessions), and nano/MATH does: on the $Q = 367$ pool, over a median per-question span of 28 minutes, $\Delta\sigma_{\mathrm{MV}}(32) = -0.031$ ($0.669 \to 0.638$) against a null s.d. of 0.055 ($p = 0.58$), and on the reference-90 pool, whose session ran 12 hours per question rather than 28 minutes, a substantial fraction of the 63-hour gap, -0.080 against 0.110 ($p = 0.47$). The between-session contrast is -0.356 against a null s.d. of 0.133, 2.7 null s.d. against 0.6 and 0.7 for the two within-session tests. That is a different cell from the one the drift is claimed on, which we cannot avoid: the AIME pools hold 32 samples per session, so no within-session split at n=32 exists there. It is, however, the matched *design*. At the budgets the AIME pools do admit, the within-session movements are unresolved and *not* small: for $\Delta\sigma_{\mathrm{MV}}(8)$ on the three nano/AIME pools (per-question spans of 4–9 minutes) they are -0.061 ($p = 0.41$), $+0.132$ ($p = 0.17$) and $+0.093$ ($p = 0.25$), two of them larger in magnitude than the between-session contrast at that same budget ($\Delta\sigma_{\mathrm{MV}}(8) = -0.073$, $p = 0.21$). We therefore do not claim within-session variation is small at n=8 on AIME; with 20–35 solvable questions nothing there resolves either way. The claim is narrower: at the budget where the between-session contrast lives, the matched within-session test is null where the between-session test is not, and the two pools' efficiency curves separate as the budget grows rather than starting apart. *The reasoning cells.* These carry the caveat in its extreme form: at the 4000 cap the truncation-matched floor sits on top of coverage ($H_{\mathrm{par}} = -0.004$ at n=16, where σ^{tm} is undefined, and $+0.004$ at n=32, a denominator of four thousandths that carries no information), so no truncation-matched value is usable there and the $\sigma = 1.00$ of §7 is entirely answer-existence filtering; at the 12,000 cap $\sigma^{\mathrm{tm}}_{\mathrm{MV}}(32) = 0.677$ against $\sigma_{\mathrm{MV}}(32) = 0.907$, and on MATH 0.230 against 0.613.

Composition or efficiency? Write $a_q = A_{\mathrm{MV},q}(n)$, $c_q = C_q(n)$, $b_q = p_q$ and $h_q = c_q - b_q$ for question q of a pool. Then $\sigma = \sum_q w_q \sigma_q$, summing over questions with $h_q > 0$, with headroom weights $w_q = h_q / \sum_{q'} h_{q'}$ and per-question efficiencies $\sigma_q = (a_q - b_q)/h_q$, so a moved σ is either a moved efficiency on the same questions or headroom moving between questions. Between

two pools, $\Delta\sigma = \sum_q \bar{w}_q \Delta\sigma_q + \sum_q \bar{\sigma}_q \Delta w_q$ exactly, with midpoint weights $\bar{w}_q$ and $\bar{\sigma}_q$; a question with headroom in only one pool has no σ_q in the other, and we set it to the observed one there ($\Delta\sigma_q = 0$), which books the question's entry or exit to composition. Intervals are question-level bootstrap. Read at n=16, where $N \geq 2n$ and coverage is not the degenerate $\mathbf{1}[K \geq 1]$ of $n = N$, the drift contrast is within-question: -0.081 $[-0.161, -0.003]$ of $\Delta\sigma_{\mathrm{MV}}(16) = -0.174$, against an unresolved composition part of -0.094 $[-0.239, +0.014]$; at n=32 the two halves are -0.178 $[-0.366, 0.000]$ and -0.177 $[-0.375, -0.005]$, three questions gaining headroom and none losing it. The cap contrast is mostly composition by point estimate: raising the cap to 3500 brings 13 questions into headroom (one leaves), and the composition part is -0.238 $[-0.398, -0.093]$ at n=16 and -0.183 $[-0.390, -0.003]$ at n=32, while the within-question part is unresolved at both (-0.031 $[-0.165, +0.083]$ and -0.081 $[-0.282, +0.093]$). The composition part includes the reweighting of questions present in both pools, not only the entrants. The longer cap lets hard questions be answered at all, and their pluralities are mostly wrong; whether voting also got worse on the questions that already had headroom, the data do not resolve.

E.5 The trained verifier

The process reward model. Qwen2.5-Math-PRM-7B in bf16 on one RTX 4090-class GPU, run with its model card's code under `transformers` 4.53.3 (PyTorch 2.13, CUDA 13.0, SDPA attention): the card's chat template and system prompt, solutions split into steps on blank lines (on single newlines when that yields a single step), the input read to the model's 4096-token context, and the per-step rewards read at the `<extra_0>` positions (a solution with no separator inside the context scores 0). We report the min over steps as the solution score, the standard choice; last and mean ship with the scores. PRM-weighted voting sums the solution scores of each answer cluster and returns the heaviest cluster. The environment has to be recorded. The card's code runs unmodified under `transformers` 4.46.3, 4.51.3 and 4.53.3 and fails under 4.54.1, 4.55.4, 4.56.2, 4.57.6 and 5.17.0 (a removed cache method, then a changed configuration class); where it runs, the card's worked example gives $[1.00, 0.16, 0.98, 1.00]$ against the published $[1.0, 0.19, 0.98, 1.0]$, and switching SDPA for eager attention moves its second step to 0.17. A second scoring of the same pools in an environment whose library version was not recorded, with inputs cut at 2048 tokens, agrees at Spearman 0.9999 on the 54,154 solutions both scored (median $|\Delta| = 0.0009$, 99th percentile 0.022); re-deriving §7.1 from it moves no σ, trend or contrast on nano, Llama or Phi by more than 0.02, the ESS rank correlation reads -0.78 instead of -0.76, and the one statistic whose resolution differs is Phi's $(4, 16)$ family contrast, whose interval's lower end is within 0.002 of zero under both scorings ($[+0.001, +0.058]$ there). The 2048-token cut touches 0.3% of MATH solutions and changes no MATH statistic, but 75% of nano's AIME solutions: under it the within-question AUC on nano/AIME falls from 0.71 and 0.63 to 0.67 and 0.53 at the 2000 and 3500 caps.

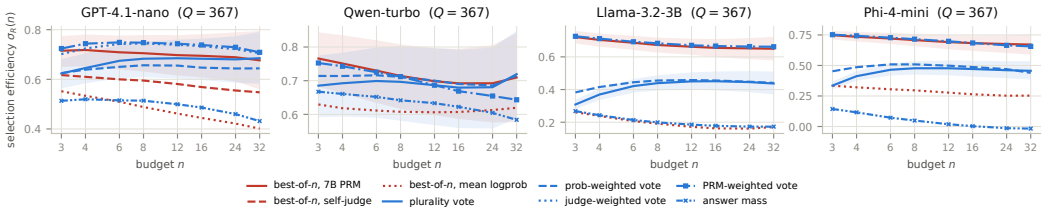


Figure 5: Selection efficiency against budget for three pointwise selectors (red) and five votes (blue), on the cells with PRM scores (§7.1): every rule of Table 6, so the votes are not a friendlier subset than the table's. Accuracy rises over this range for every rule but answer mass on Phi, which falls below the content-blind floor; the efficiency axis shows the two families moving apart. The two marked votes are the concentrated ones (§7.1): answer mass on every cell and PRM-weighted voting on Llama and Phi (it is marked on every panel for comparison; on nano its weights are spread and it does not fall, and on Qwen it falls although they are spread). Bands are 95% question-level bootstrap for best-of-n under the PRM and for plurality voting. n=2 is omitted: there a plurality is a coin flip whenever the two answers differ ($\sigma_{\mathrm{MV}}(2) = 0$ exactly) and a weighted vote reduces to best-of-2, so that budget mixes two unrelated effects. The judge rules exist only on the nano pool.

Table 5: The paired family contrast of §7.1 and the two family means it is built from, per cell, with question-level bootstrap intervals (* excludes zero). Positive means efficiency *falls*. The contrast is the selector mean minus the vote mean; the last column is plurality voting's own trend, which is the vote the selector-versus-vote framing is usually about. The last three columns are at $(4, 16)$. The selector means fall resolvedly on every cell but Qwen; the vote means are unresolved at this window, though individual votes move in both directions (Table 6); plurality voting's own efficiency rises on nano, Llama and Phi, resolvedly on the last two, and is flat on Qwen ($+0.013$). Accuracy rises over the window for every rule but answer mass on Phi (nano: 0.760→0.793 for the PRM, 0.753→0.791 for voting).

Cell	contrast $(4, 16)$	contrast $(3, 15)$	sel. mean	vote mean	MV alone
nano	$+0.054$ $[+0.018, +0.090]^*$	$+0.079$ $[+0.041, +0.117]^*$	$+0.049^*$	-0.005	-0.037
Qwen	$+0.002$ $[-0.044, +0.049]$	$+0.015$ $[-0.036, +0.068]$	$+0.039$	$+0.037$	$+0.013$
Llama	$+0.065$ $[+0.038, +0.092]^*$	$+0.093$ $[+0.064, +0.123]^*$	$+0.063^*$	-0.003	-0.081^*
Phi	$+0.028$ $[0.000, +0.056]$	$+0.057$ $[+0.027, +0.087]^*$	$+0.055^*$	$+0.027$	-0.059^*

Robustness of §7.1. A weighted vote is a vote only insofar as its weights are spread, which the effective sample size $\mathrm{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$ measures: answer mass weights each sample by its sequence probability $\exp(\sum_t \log p_t)$, leaving $\mathrm{ESS} \leq 0.18n$ at n=16 and making its winner the top-weight sample's own answer on ≥ 0.997 of draws, and PRM-weighted voting concentrates, less strongly, on Llama ($0.55n$) and Phi ($0.64n$), where the PRM discriminates best. Over all 17 (cell, vote) pairs at n=16, ESS is 0.07–0.17n for answer mass, 0.55–0.93n for PRM-weighted, 0.86–1.00n for probability-weighted and $0.92n$ for judge-weighted voting, and n for plurality. Benjamini–Hochberg (Benjamini & Hochberg, 1995) at $q = 0.05$ across the 38 tests behind §7.1's trend claims (the 26 per-rule $(3, 16)$ trends and 12 paired contrasts) leaves 17 of the 19 nominal exclusions standing; the two it drops are the PRM's and PRM-weighted voting's trends on Qwen ($p = 0.028$ and 0.049), and the odd-budget window $(3, 15)$ still resolves on three cells and $(4, 16)$ on two. §7.1's other intervals are uncorrected. The PRM was trained on MATH-family problems (Zhang et al., 2025); on AIME, where all four cells are scored ($Q = 60$), the $(3, 16)$ trends of best-of-n under the PRM are -0.015, $+0.146^*$, 0.000, $+0.205^*$ on nano at the 2000 and 3500 caps, Phi and Llama (* resolved), two positive and two indistinguishable from zero, and the mean-logprob selector's are $+0.051$, $+0.121$, $+0.213^*$, $+0.117$. With the PRM's input cut at 2048 tokens, which truncates 75% of nano's AIME solutions, its trends become $+0.028$, $+0.119^*$, $+0.006$, $+0.231^*$: the two that resolve do so under both readings, and the other two stay indistinguishable from zero. Robustness of the paired contrast of §7.1: dropping each rule from either family in turn never changes its sign on nano and Llama or on Phi at $(3, 15)$, and on nano, Llama and Phi it still excludes zero in 34 of the 40 re-runs over the two windows; read over $(4, 32)$ it keeps its sign on nano, Llama and Phi and excludes zero on Llama, and on Qwen it turns negative, unresolved. Against the truncation-matched floor the paired contrast keeps its sign on nano, Llama and Phi and still excludes zero at $(3, 15)$ on all three; on Qwen that floor exceeds coverage at n=3, so σ^{tm} is undefined there.

What the normalization buys. $\sigma_R(n)$ is an affine rescaling of $A_R(n)$ at a fixed cell and budget, so the paired family contrast of §7.1 was also run on the raw accuracy gap $A_{\text{sel}} - A_{\text{vote}}$, on the identical question resamples, at $(4, 16)$: nano $+0.054$ $[+0.018, +0.090]$ against $+0.0099$ $[+0.0043, +0.0155]$; Qwen $+0.002$ $[-0.044, +0.049]$ against $+0.0007$ $[-0.0047, +0.0061]$; Llama $+0.065$ $[+0.038, +0.092]$ against $+0.0188$ $[+0.0068, +0.0306]$; Phi $+0.028$ $[0.000, +0.056]$ against -0.0019 $[-0.0120, +0.0084]$. It resolves on nano and Llama, where the efficiency contrast does, and not on Qwen or Phi, where the efficiency contrast is null or at the boundary; on Phi it has the opposite sign. The two contrasts are not the same statistic: the efficiency contrast is $g(4)/H(4) - g(16)/H(16)$ for an accuracy lead g of the selectors over the votes, and since headroom grows with n an unchanged lead registers as a falling efficiency contrast. On Phi that is all of it: the lead is unchanged, and the efficiency contrast's point estimate is positive only because headroom grows by 60% between n=4 and n=16. On nano and Llama the accuracy contrast resolves on its own.

A second verifier. Math-Shepherd (Wang et al., 2024a), a Mistral-7B PRM (none of the four scored samplers is from its family), scored the same four MATH cells in its card's format: each step written *Step k:* … and closed by the card's step tag, the step reward $P(+)/(P(+) + P(-))$ read at the tag, the solution score the minimum over steps, the input read to 4096 tokens, bf16 under the same

Table 6: Selection efficiency $\sigma_R(n)$ for every rule and budget behind §7.1, on the four MATH cells with complete PRM scores. Selectors are above the horizontal line, votes below; the judge rules exist only where the judge scored the pool. Each cell is restricted to the 32 PRM-scored samples a question has, so $N = 32$ throughout, the n=32 column is the whole pool (one subset, no averaging), and these σ values differ from Tables 1 and 2, which use the full N. All four cells are the same 367 MATH questions, every sample scored by the PRM. The last two columns are the budget trend $\sigma_R(a) - \sigma_R(b)$ with its question-level bootstrap interval, over the window $(3, 16)$ that §7.1 quotes, and over $(3, 32)$ (point estimate, * marking exclusion of zero; its n=32 end is the whole pool); positive means the rule converts a shrinking share of the headroom as the budget grows, and the prose counts signs over exactly these rows. No selector trend runs the other way; among votes, answer mass on all four cells, plurality and probability-weighted voting on Qwen, and PRM-weighted voting on Qwen, Llama, and Phi do.

Cell	Rule	n=3	n=4	n=8	n=15	n=16	n=24	n=32	$\sigma(3)-\sigma(16)$	$\sigma(3)-\sigma(32)$
nano	best-of-n under the PRM	0.723	0.715	0.703	0.697	0.697	0.690	0.676	$+0.026$ $[-0.034, +0.088]$	$+0.046$
	self-judge argmax	0.615	0.609	0.595	0.573	0.571	0.556	0.550	$+0.044$ $[-0.003, +0.093]$	$+0.065$
	max mean-token-logprob	0.554	0.535	0.486	0.448	0.444	0.420	0.401	$+0.110$ $[+0.040, +0.177]^*$	$+0.153^*$
	plurality vote	0.624	0.646	0.683	0.684	0.683	0.680	0.685	-0.059 $[-0.127, +0.007]$	-0.062
	probability-weighted vote	0.628	0.637	0.654	0.650	0.648	0.643	0.644	-0.020 $[-0.092, +0.049]$	-0.016
	judge-weighted vote	0.705	0.725	0.744	0.738	0.734	0.722	0.706	-0.029 $[-0.094, +0.037]$	-0.001
	PRM-weighted vote	0.723	0.740	0.749	0.740	0.739	0.729	0.709	-0.017 $[-0.077, +0.046]$	$+0.014$
	answer mass	0.512	0.517	0.512	0.488	0.485	0.459	0.431	$+0.026$ $[-0.039, +0.090]$	$+0.081$
Qwen	best-of-n under the PRM	0.767	0.757	0.716	0.691	0.692	0.692	0.710	$+0.075$ $[+0.007, +0.145]^*$	$+0.057$
	max mean-token-logprob	0.624	0.617	0.607	0.604	0.605	0.612	0.620	$+0.019$ $[-0.068, +0.106]$	$+0.004$
	plurality vote	0.685	0.692	0.697	0.681	0.679	0.681	0.719	$+0.006$ $[-0.071, +0.079]$	-0.034
	probability-weighted vote	0.708	0.709	0.712	0.691	0.689	0.687	0.715	$+0.018$ $[-0.061, +0.096]$	-0.007
	PRM-weighted vote	0.744	0.747	0.711	0.671	0.667	0.655	0.643	$+0.077$ $[0.000, +0.152]^*$	$+0.100$
	answer mass	0.658	0.661	0.645	0.625	0.625	0.603	0.584	$+0.033$ $[-0.056, +0.124]$	$+0.075$
Llama	best-of-n under the PRM	0.722	0.705	0.672	0.657	0.655	0.653	0.650	$+0.066$ $[+0.024, +0.109]^*$	$+0.072^*$
	max mean-token-logprob	0.258	0.239	0.190	0.166	0.164	0.161	0.171	$+0.094$ $[+0.054, +0.133]^*$	$+0.087^*$
	plurality vote	0.309	0.370	0.439	0.451	0.451	0.445	0.437	-0.142 $[-0.192, -0.094]^*$	-0.128^*
	probability-weighted vote	0.382	0.418	0.454	0.456	0.455	0.449	0.441	-0.073 $[-0.124, -0.022]^*$	-0.058
	PRM-weighted vote	0.728	0.708	0.681	0.668	0.667	0.664	0.662	$+0.062$ $[+0.019, +0.104]^*$	$+0.066^*$
	answer mass	0.272	0.243	0.199	0.178	0.177	0.172	0.173	$+0.096$ $[+0.058, +0.133]^*$	$+0.099^*$
Phi	best-of-n under the PRM	0.745	0.735	0.708	0.685	0.685	0.675	0.671	$+0.060$ $[+0.015, +0.105]^*$	$+0.074^*$
	max mean-token-logprob	0.332	0.323	0.295	0.266	0.263	0.252	0.253	$+0.069$ $[+0.024, +0.113]^*$	$+0.079$
	plurality vote	0.335	0.412	0.477	0.473	0.471	0.461	0.456	-0.136 $[-0.189, -0.084]^*$	-0.121^*
	probability-weighted vote	0.455	0.487	0.508	0.490	0.487	0.471	0.440	-0.032 $[-0.087, +0.021]$	$+0.015$
	PRM-weighted vote	0.748	0.742	0.715	0.688	0.685	0.664	0.656	$+0.064$ $[+0.016, +0.111]^*$	$+0.092^*$
	answer mass	0.143	0.114	0.049	0.004	0.003	-0.015	-0.017	$+0.139$ $[+0.090, +0.189]^*$	$+0.160^*$

`transformers` 4.53.3 (it reproduces the card's worked example to within 0.002). It ranks less well, within-question AUC 0.63, 0.64, 0.66, 0.72 on nano, Qwen, Llama and Phi against the Qwen PRM's 0.76, 0.72, 0.82, 0.86, and best-of-n under it has a positive $(3, 16)$ trend on all four cells ($+0.063^*$, $+0.045$, $+0.130^*$, $+0.116^*$), resolved on nano, Llama and Phi and larger than the Qwen PRM's on nano, Llama and Phi. With Math-Shepherd-weighted voting in place of PRM-weighted voting, the paired family contrast excludes zero on nano, Llama and Phi at both $(4, 16)$ and $(3, 15)$, and the ESS of the 17 (cell, vote) pairs ranks their $(3, 16)$ trends at Spearman -0.74 ($p = 0.0007$). Both PRMs were trained on MATH-family problems (Zhang et al., 2025; Wang et al., 2024a), so neither is an out-of-distribution test; what the second one shows is that the selector decay is not a property of a single verifier.

Fresh pools. The Qwen PRM also scored the fresh pools of Appendix E.4, two sessions each of Phi and Llama (Phi A, Phi B, Llama A, Llama B; within-question AUC 0.86, 0.86, 0.79, 0.79, against 0.86 and 0.82 on the reference Phi and Llama pools). The paired family contrast of §7.1 holds on them, excluding zero on all four pools at $(4, 16)$ ($+0.042$, $+0.035$, $+0.062$, $+0.058$) and on all four at $(3, 15)$, and, as on the reference cells, on none at $(8, 16)$ ($+0.008$, 0.000, $+0.016$, $+0.015$); the PRM's own $(3, 16)$ trend is positive on all four and resolves on three; and the ESS of the weights ranks the $(3, 16)$ trends of the four votes (plurality, probability-weighted, PRM-weighted, answer mass) at Spearman -0.95 over Phi's eight (session, vote) pairs and -0.85 over Llama's, each family's two sessions counting as two observations of the same four rules. Math-Shepherd on the same pools agrees: its best-of-n trend is positive on all four (all resolving) and the

family contrast excludes zero at $(4, 16)$ on all four. These contrasts and the confidence-filtered ones below are outside the Benjamini–Hochberg family of §7.1 and are uncorrected.

A larger model. To see whether the verifier findings hold on a larger model that both PRMs score, Llama-3.1-8B-Instruct (Grattafiori et al., 2024) was served for one session on the stack and settings of the fresh pools (vLLM 0.30.0, a 4096-token window), 32 samples on each of the 367 MATH questions, and both PRMs scored every sample. At the 1400-token cap it truncates 16.2% of its generations (Llama-3.2-3B: 7.8% and 7.5%); $B = 0.387$, $C(32) = 0.815$, $A_{\mathrm{MV}}(32) = 0.614$ and $\sigma_{\mathrm{MV}}(32) = 0.53$ $[0.46, 0.61]$. On this one session the verifier findings of §7.1 recur (uncorrected). The Qwen PRM ranks its samples at within-question AUC 0.78, and best-of-n under it loses efficiency as the budget grows ($(3, 16)$ trend $+0.074$ $[+0.033, +0.114]$), as do best-of-n by mean log-probability ($+0.123^*$) and the most confident trace (DeepConf's lowest-group confidence, next paragraph; $+0.060^*$); among votes, plurality gains (-0.146^*), while answer mass ($+0.095^*$) and PRM-weighted voting, whose weights concentrate (ESS $0.58n$ at n=16), decay ($+0.053^*$), as on Llama-3.2-3B and Phi. The paired family contrast is $+0.082$ $[+0.052, +0.113]$ at $(4, 16)$, $+0.115$ $[+0.083, +0.147]$ at $(3, 15)$ and $+0.027$ $[+0.009, +0.046]$ at $(8, 16)$, where ties are rare, and the ESS of the weights orders the $(3, 16)$ trends of the four votes exactly (Spearman -1.00 over four rules). Math-Shepherd (AUC 0.68) agrees: best-of-n under it has trend $+0.118^*$ and the family contrast at $(4, 16)$ is $+0.111^*$.

Confidence-filtered voting. The account of §7.1 says a vote whose weights concentrate decays like a selector. Offline DeepConf-style filtering (Fu et al., 2025) concentrates them by construction, so we evaluated it on the fresh pools of Appendix E.4 (Phi A, Phi B, Llama A, Llama B), whose requests also returned the top-10 log-probabilities of every generated token (a response field that does not affect sampling). The token confidence is $c_t = -\frac{1}{10}\sum_{j=1}^{10} \log p_t(j)$, $p_t(j)$ the j-th largest next-token probability at position t, and a trace's confidence is its lowest mean c_t over a sliding window of 64 tokens (DeepConf's lowest-group measure, with the window scaled to these few-hundred-token traces rather than its default 2048; within-question AUC 0.60–0.66). The rule keeps the $\lceil \eta n \rceil$ most confident of n traces and takes a confidence-weighted plurality among them (Table 7). At $\eta = 0.1$, DeepConf's aggressive setting, which for $n \le 20$ keeps one or two traces and is therefore close to the most-confident-trace selector, the $(3, 16)$ trend is positive on all four pools ($+0.030$, $+0.029$, $+0.046^*$, $+0.013$), a selector's sign; at $\eta = 0.9$ it is negative on all four (-0.083^*, -0.085^*, -0.022, -0.046^*), a vote's. At $\eta = 0.25$ and 0.5 the filtered vote gains about as plurality does ($(3, 16)$: -0.125^*, -0.118^*, -0.087^*, -0.116^* and -0.189^*, -0.185^*, -0.112^*, -0.138^*; plurality -0.176^*, -0.174^*, -0.103^*, -0.123^*), and not only because it keeps a single trace at n=3: between n=8 and n=16, where it keeps two and four ($\eta = 0.25$) or four and eight ($\eta = 0.5$) traces, it still gains on every pool (-0.140^*, -0.123^*, -0.093^*, -0.106^* and -0.043^*, -0.048^*, -0.030^*, -0.045^*; plurality -0.008, -0.004, $+0.011$, -0.005). Its ESS is about $\lceil \eta n \rceil$, a fraction of n that moves with n (Table 7); at n=16 it is about ηn, so the cut ESS $< 0.65n$ predicts a fall at $\eta = 0.25$ and 0.5 and is wrong there, and the ESS ordering is not monotone ($\eta = 0.5$ gains more than $\eta = 0.9$). The ESS account describes votes that weight each sample by a fixed function of that sample, and a filter that keeps $\lceil \eta n \rceil$ of them is outside it. Table 7 shows the difference: answer mass, which decays, keeps about one effective sample at every budget (ESS $0.35n$ at n=3, $0.08n$ at n=16), whereas the filtered votes at $\eta \ge 0.25$ keep a number that grows with n. The count of effective samples is not the whole story either: PRM-weighted voting on Llama and Phi keeps $0.55n$ and $0.64n$ at n=16, more than the $\eta = 0.5$ filter's $0.50n$, and decays where the filter gains (§7.1). DeepConf's other trace confidences (bottom-10% group, tail, mean token confidence) give the same picture: at $\eta = 0.1$ the trend is positive on 11 of 12 (pool, measure) pairs, at $\eta \ge 0.25$ negative on 48 of 48.

E.6 Beyond mathematics

Every cell above is competition mathematics. To see whether the instrument carries over, Phi-4-mini and Llama-3.2-3B answered two other tasks, served as the fresh pools of Appendix E.4 (32 samples per question at $T = 0.8$, the 1400-token cap). *MMLU-Pro* (Wang et al., 2024b): 360 questions, 40 from each of the nine categories outside mathematics, the physical sciences, engineering and computer science (law, history, philosophy, psychology, health, biology, economics, business, other), every one with ten options, under MMLU-Pro's zero-shot chain-of-thought instruction; the answer is the letter after the last "answer is". *TriviaQA* (Joshi et al., 2017): 360 questions drawn at random

Table 7: Offline confidence-filtered voting on the fresh open-weight pools (A and B: two sessions of the same model, serving stack and settings): the $(3, 16)$ efficiency trend $\sigma_R(3) - \sigma_R(16)$ (positive means the rule converts a shrinking share of the headroom; * question-level bootstrap interval excludes zero) and the ESS of each rule's weights as a fraction of n at n=3 and n=16, averaged over the pools. Trace confidence is the lowest-group measure; the other measures give the same signs (text).

Rule	ESS/n, n=3	n=16	Phi A	Phi B	Llama A	Llama B
most confident trace	0.33	0.06	$+0.047$	$+0.041$	$+0.074^*$	$+0.034$
top 10%, weighted vote	0.33	0.12	$+0.030$	$+0.029$	$+0.046^*$	$+0.013$
top 25%, weighted vote	0.33	0.25	-0.125^*	-0.118^*	-0.087^*	-0.116^*
top 50%, weighted vote	0.66	0.50	-0.189^*	-0.185^*	-0.112^*	-0.138^*
top 75%, weighted vote	0.99	0.74	-0.087^*	-0.079^*	-0.008	-0.037
top 90%, weighted vote	0.99	0.92	-0.083^*	-0.085^*	-0.022	-0.046^*
plurality vote	1.00	1.00	-0.176^*	-0.174^*	-0.103^*	-0.123^*
answer mass	0.35	0.08	$+0.161^*$	$+0.118^*$	$+0.082^*$	$+0.114^*$
max mean-logprob	0.33	0.06	$+0.095^*$	$+0.052^*$	$+0.115^*$	$+0.094^*$

from the validation split of its no-context reading-comprehension set, prompted to end with a line "Answer: ..." and graded by exact match against TriviaQA's normalized aliases, the benchmark's own metric; other answers cluster by normalized string. A response without an answer abstains. Table 8 gives the decomposition and the trends.

Multiple choice. The decomposition applies unchanged, and for the same two models it reports a lower efficiency than on MATH. The headroom at n=16, where the coverage estimate is not degenerate, is what it is on MATH on the same serving stack (Phi 0.309 against 0.307 and 0.323; Llama 0.387 against 0.390 and 0.384 in the two fresh MATH sessions), but plurality voting converts far less of it: $\sigma_{\mathrm{MV}}(16)$ is 0.21 against 0.50 and 0.50 for Phi, 0.20 against 0.42 and 0.44 for Llama. Nor does it rise with the budget as it does on MATH: from n=3 to n=16, σ_{MV} goes $0.20 \to 0.21$ and $0.18 \to 0.20$ (the $(3, 16)$ trends -0.010 $[-0.050, +0.027]$ and -0.021 $[-0.057, +0.012]$, unresolved, against -0.176^* and -0.174^*; -0.103^* and -0.123^* on MATH). The samples concentrate on few letters (a median of 3 and 4 distinct answers per question, against 6–7 and 10–11 on MATH), so wrong samples coincide: the correct answer is the strict plurality of the 32 on 67% and 59% of solvable questions against 79%–80% and 64%–67%. The mode is settled early, and further samples add coverage that no vote converts. Part of that coverage is guessing: with a handful of candidate letters, a correct letter somewhere among 32 samples need not reflect a solution, so on multiple choice $C(n)$, and with it H, overstates what the model can solve, and a low σ partly records a vote that declines to follow a guess. Guessing raises B too, so the decomposition does not remove it; it only shows where it sits.

Free-form answers. TriviaQA behaves like MATH (Phi, then Llama). Plurality converts $\sigma_{\mathrm{MV}}(16) = 0.38$ and 0.42 of the headroom ($H(16) = 0.273$ and 0.210) and gains with the budget, from 0.26 and 0.29 at n=3 ($(3, 16)$ trends -0.119 $[-0.177, -0.063]$ and -0.124 $[-0.186, -0.062]$), as it does on MATH; the correct answer is the strict plurality of the 32 on 69% and 80% of solvable questions. Exact match is strict ("freshwater fish" for *fish* is wrong), which moves correct paraphrases into wrong clusters and lowers B and C together.

The rules (values for MMLU-Pro Phi, MMLU-Pro Llama, TriviaQA Phi and TriviaQA Llama, in that order). Pointwise selectors lose efficiency on all eight (task, model, selector) triples, four of them resolved: best-of-n by mean log-probability $+0.025$, $+0.058^*$, $+0.073^*$ and $+0.024$, the most confident trace $+0.016$, $+0.083^*$, $+0.053^*$ and $+0.040$; DeepConf's top-10% vote, close to a selector at these budgets, likewise ($+0.010$, $+0.069^*$, $+0.056^*$ and $+0.036$). Answer mass decays on multiple choice ($+0.062^*$ and $+0.056^*$) and does not resolve a trend on TriviaQA ($+0.045$ and -0.014). On multiple-choice Llama the most confident trace ends at $\sigma = -0.029$ at n=32 (point estimate), no better than a random pick. Among votes, plurality gains on TriviaQA and not on multiple choice, as above, and the probability-weighted, top-75% and top-90% votes follow it without resolving ($+0.009$, -0.006, -0.030 and -0.043; $+0.013$, $+0.005$, -0.031 and -0.036; $+0.006$, -0.001, -0.050 and -0.054). The confidence-filtered votes at $\eta = 0.25$ and 0.5 gain on every pool (-0.068^*, -0.037^*, -0.079^* and -0.059^*; -0.095^*, -0.080^*, -0.160^* and -0.165^*), as on MATH (Table 7), by keeping one or two traces at small n and more as n grows. The trace

confidence separates correct from wrong samples of the same question less well than on MATH (AUC 0.54, 0.55, 0.59 and 0.63, against 0.66 and 0.61 for Phi and Llama on MATH).

Table 8: MMLU-Pro and TriviaQA ($Q = 360$ each, $N = 32$). Top: the decomposition at $n = 32$ as in Table 1 (95% question-level bootstrap; with $n = N$, $C(32) = Q_s/Q$). Bottom: $(3, 16)$ efficiency trends $\sigma_R(3) - \sigma_R(16)$ (positive means the efficiency falls; * the interval excludes zero); the rules are those of Table 7.

	MMLU-Pro		TriviaQA	
	Phi	Llama	Phi	Llama
Q_s	313	317	265	279
B	0.522	0.432	0.411	0.532
$C(32)$	0.869	0.881	0.736	0.775
$H(32)$	0.348	0.448	0.325	0.243
$A_{\mathrm{MV}}(32)$	0.594	0.529	0.518	0.624
Gap	0.275 [0.231, 0.321]	0.351 [0.301, 0.397]	0.219 [0.177, 0.261]	0.151 [0.117, 0.190]
$\sigma_{\mathrm{MV}}(32)$	0.21 [0.14, 0.28]	0.22 [0.16, 0.28]	0.33 [0.25, 0.41]	0.38 [0.28, 0.47]
wrong-mode share	0.31	0.40	0.29	0.19
abstention	0.019	0.039	0.000	0.019
max mean-logprob	$+0.025$	$+0.058^*$	$+0.073^*$	$+0.024$
most confident trace	$+0.016$	$+0.083^*$	$+0.053^*$	$+0.040$
top 10%, weighted vote	$+0.010$	$+0.069^*$	$+0.056^*$	$+0.036$
top 25%, weighted vote	-0.068^*	-0.037^*	-0.079^*	-0.059^*
top 50%, weighted vote	-0.095^*	-0.080^*	-0.160^*	-0.165^*
top 75%, weighted vote	$+0.013$	$+0.005$	-0.031	-0.036
top 90%, weighted vote	$+0.006$	-0.001	-0.050	-0.054
plurality vote	-0.010	-0.021	-0.119^*	-0.124^*
prob-weighted vote	$+0.009$	-0.006	-0.030	-0.043
answer mass	$+0.062^*$	$+0.056^*$	$+0.045$	-0.014

E.7 Adaptive schedules

The adaptive channels carry question-level intervals on the same resamples as the totals, and the paired difference (selection minus headroom) excludes zero on none of the three cells: nano -0.004 $[-0.026, +0.015]$, Llama $+0.009$ $[-0.002, +0.022]$, Phi $+0.004$ $[-0.012, +0.022]$. By point estimate the headroom channel is negative for one schedule or both on 6 of the nine (Llama/AIME, the original 2000-cap pool, the 2000-cap rerun, nano/AIME at the matched 3500 cap, Phi/AIME, Mistral/MATH), in every case where coverage at the schedule's own budget is below 0.55. On Llama/AIME, Phi/AIME and the matched 3500 cap it is within 0.001 of zero, and it resolves only for early stopping on the original 2000-cap pool (-0.019 $[-0.033, -0.007]$). That stratification is not question count in disguise: Mistral/MATH carries the same $Q = 90$ as the three cells that gain and does not gain, and subsampling each of those three to $Q = 60$, the AIME cells' size, leaves the contrast resolved in 90–98% of draws. The 0.55 threshold is nonetheless read off these nine cells, and nine cells cannot support an interaction test. The schedules run at their authors' defaults: adaptive-consistency at a Dirichlet stopping threshold of 0.95 with $n_{\min}$=2, early stopping with a window of 5. One cross-cell regularity does not survive: the saving tracks coverage more than efficiency for adaptive-consistency ($r = 0.82$ vs. 0.35), but the ordering reverses for early stopping (0.48 vs. 0.60) and the contrast is unresolved at 9 cells $[-0.068, +1.071]$, so we do not claim it. *The adaptive schedules, in detail.* A schedule's expected budget is not an integer, so the fixed-n baseline it is compared against is read off the fixed-budget curve by linear interpolation between $\lfloor \mathbb{E}[n] \rfloor$ and $\lceil \mathbb{E}[n] \rceil$. The interpolation is the accuracy of a randomized fixed-budget policy that spends the same expected budget, so it is the matched baseline; the conservative check compares instead against the integer budget $\lceil \mathbb{E}[n] \rceil$, which spends at least as many samples as the schedule does on average: adaptive-consistency's advantage is then $+0.020$ nano, $+0.014$ Llama, $+0.017$ Phi, all three excluding zero, so the interpolation is not what produces it (early stopping then resolves on nano only). Each schedule is simulated along 200 random orderings of every 32-sample pool; the budget is a stopping time, so for each ordering the schedule's accuracy and coverage are those of the samples it actually drew, the fixed-budget baseline's are read at its expected budget as above, the identity of Theorem 1 is applied to that pair, and the channels are averaged over orderings. Lemma 1's calibration holds only at a fixed n, so these channels are an accounting of the matched contrast, not efficiencies on the $[0, 1]$ scale.

E.8 BIAS OF THE RATIO, AND CELLS WITHOUT A FINITE INTERVAL

σ is a ratio of unbiased estimates, so the plug-in $\hat{\sigma}$ is not itself unbiased, and a percentile interval is only as good as the resamples behind it. We therefore report both diagnostics for *every* row of Table 1 rather than for a convenient subset. Jackknifing over questions puts the bias at 0.0000 to 0.0059 across all 14 cells (largest on GPT-4.1-nano/AIME 2000 rerun), one to two orders of magnitude below the drift and cap contrasts, so we quote the plug-in value; per-cell values ship with the results. The interval is the more delicate object: a question resample with $\hat{C} = \hat{B}$ makes $\hat{\sigma}$ a $0/0$, and our percentile interval drops those resamples, which conditions it on non-degeneracy. On 13 of the 14 cells this never happens in 8000 resamples. It happens on one, Mistral-7B/AIME with $Q_{\mathrm{s}} = 3$ of 60, where 409 resamples (5.1%) are degenerate, and the Fieller interval, which is the honest construction when a denominator can reach zero, is *unbounded*: the data cannot rule out $H = 0$, so no finite interval for σ exists there and Table 1 prints none. Elsewhere the two constructions agree to within 0.009 on the four $Q = 367$ cells, and every divergence above 0.05, up to a largest of 0.11, is a cell whose percentile interval is truncated at $\sigma = 1$ while the Fieller set (Fieller, 1954) extends past it; since $\sigma \leq 1$, that divergence is a property of the constructions rather than of the data. This is the quantitative form of the warning in §8 that a cell with little headroom cannot adjudicate anything.