YOU CANNOT VOTE YOUR WAY TO THE TRUTH: THE RANK LAW OF VERIFICATION-FREE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated sampling is a standard way to spend inference-time compute: coverage, the probability that at least one of k samples is correct, can keep increasing with k. Yet a deployed system must return one answer without a gold label, and single-answer accuracy often saturates after only a few dozen samples. We characterize this *selection ceiling* by a single integer for each problem, R^*: the mass rank of the correct answer in the model's per-problem output distribution. The count-ranked shortlist frontier is then a functional of the distribution of R^*: self-consistency converges to $\mathbb{P}[R^*{=}1]$, a top-L shortlist to $\mathbb{P}[R^*{\le}L]$, and coverage to $\mathbb{P}[R^*{<}\infty]$, with exponential finite-k convergence. This turns the gap into an information price: an ideal target-aware channel with M messages is worth exactly the Bayes top-M posterior mass, the count-ranked shortlist is a measurable realization of that capacity and coincides with it under rank calibration, so resolving a top-L count-ranked ambiguity costs $\log_2 L$ oracle bits, which a real verifier need not collect. Across 10 model–task pairs, three model families, and budgets to $k{=}10^4$, R^* masses estimated from half of each problem's samples predict held-out selection accuracy across budgets and shortlist depths with median absolute error 0.002 and no fitted parameter. Across 12 bridgeable difficulty strata, a count-ranked four-answer shortlist plus two such ideal bits would recover 77–100% of the gap because, when plurality fails, the truth is usually the runner-up. Yet extracting those bits from the model is difficult: no count-based selector we test beats plurality, and only 1 of 26 implementable verifier conditions does; even there, spending the same tokens on a stronger generator performs better. Repeated sampling can therefore make the truth *available without making it identifiable*. The selection gap is not a decoding problem; it is missing information, and it can be priced.

1 INTRODUCTION

Repeated sampling is the cheapest way to spend inference compute on a reasoning model, and for *generation* it works: coverage, the fraction of problems for which at least one of k samples is correct, grows as a power law in k over four orders of magnitude (Brown et al., 2024). Deployment needs one answer chosen without a gold label, and there the returns stop. Self-consistency (Wang et al., 2023) saturates within a few dozen draws while coverage keeps climbing, and recent work has established the phenomenon carefully: plurality voting converges to a budget-independent *modal ceiling*, aggregation amplifies rather than cancels shared misconceptions, and a model scoring its own candidates recovers almost none of the gap (Bay & Yearick, 2026; Denisov-Blanch et al., 2026; Ding, 2026; Liu, 2026).

We take the phenomenon as established and ask the two questions it leaves open. *What determines this selection ceiling?* And *what would it cost to escape?* Our answer comes in two parts, and they should not be run together. The general part is a *bound*: drawing k samples is an *estimation* procedure for the model's per-problem output law, while which answer is true is a separate *identification* question fixed before any sample is drawn. Nothing a selector computes from the sample set can therefore exceed the ceiling set by that law, at any budget; and for any feature map read per sample, the ceiling tightens to the one set by the feature's own law (Theorem 2.1). The sharp part is a *characterization* of that ceiling for the feature map deployed systems actually use, the answer counts. Let θ be the model's per-problem distribution over answers and sort the alphabet by θ-mass; let

$$R^* \;=\; \text{the position of the true answer in that ordering.}$$

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.

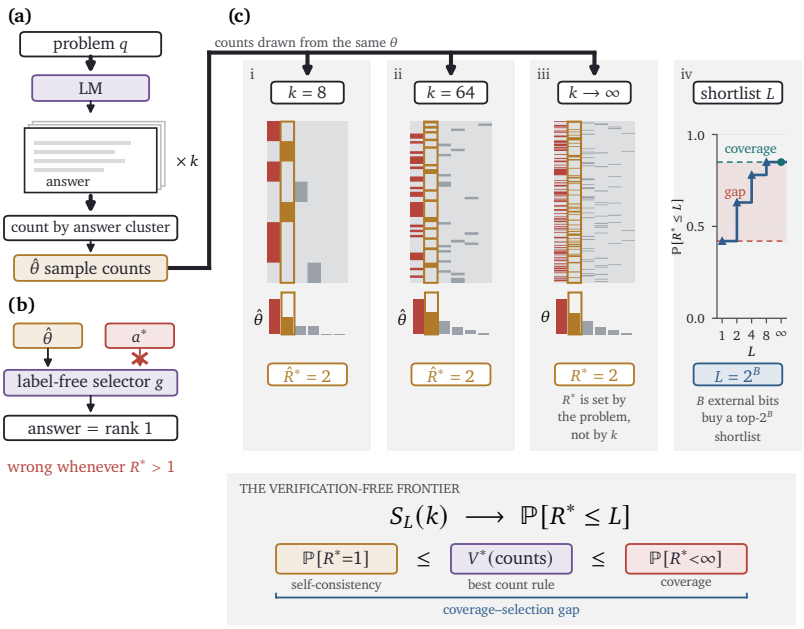


Figure 1: **Schematic; no quantity in this figure is a measurement.** **(a)** A problem is sampled k times, equivalent outputs are clustered, and the counts are $\hat{\theta}$. **(b)** What a selector may read: a label-free g sees $\hat{\theta}$ and nothing else, while a^* only ever measures ceilings, so answering the plurality is wrong on exactly the problems with $R^*>1$. **(c)** The same sample–answer incidence matrix at k=8, k=64 and in the limit **(i–iii)**: one row per sample, lit in the column of the answer it produced, columns ordered by θ-mass. Sampling *estimates*: $\hat{\theta}$ sharpens while the outlined column does not move, so R^* is a property of the problem and its fixed generation policy, not of the budget. Red is the modal answer, which voting returns and which is wrong here; gold is the truth, at R^*=2, the case this paper is about. **(iv)** The identification question no amount of sampling touches: the count-ranked shortlist frontier is one functional of R^*, so it is that integer's CDF and the shaded band is the coverage–selection gap. The depth axis reads L=2^B because B external bits buy a top-2^B *count-ranked* shortlist, worth at most the Bayes capacity G_{2^B} and equal to it under Corollary 2.4. Figure 2 is the measured version.

Then the whole *count-ranked shortlist* frontier is a functional of that one integer. It is an *information* frontier: only the first point is an achievable selector accuracy, and the rest price what external information would buy. Self-consistency tends to $\mathbb{P}[R^*=1]$, a top-L count-ranked shortlist to $\mathbb{P}[R^*\le L]$, and coverage to $\mathbb{P}[R^*<\infty]$. An idealized target-aware channel with M messages is worth exactly the Bayes top-M posterior mass G_M, and spending its $\log_2 M$ bits inside the count-ranked top-M shortlist realizes $\mathbb{P}[R^*\le M]$ of that capacity, all of it under rank calibration. Throughout, *a bit is oracle capacity, not an achievable accuracy*: what a channel that already knows the truth could supply, which §7 shows real verifiers barely collect.

That reframing changes what should be measured and what should be built. A coverage–selection gap of 0.118 sounds like a large amount of missing information; on the same cell, *one ideal bit* per problem, spent inside the count-ranked top-two shortlist, would close 73% of it. The gap does not behave like a signal-processing problem to be attacked with better label-free rules, since a learned rule far richer than the count statistics the literature proposes recovers nothing. It behaves like a small, measurable quantity of external information, and how small is the paper's most transferable number: across the 12 bridgeable difficulty strata, a count-ranked four-answer shortlist plus *two* ideal disambiguation bits would close 0.77–1.00 of it, because when the plurality is wrong the truth is usually the runner-up. The prescription, stated in full in §9, is to expose a two-element shortlist for one cheap external check where the model is moderately uncertain, and to spend on generation on the hard tail, where the truth's rank on failures runs into the tens and no small verification budget reaches it.

Contributions.

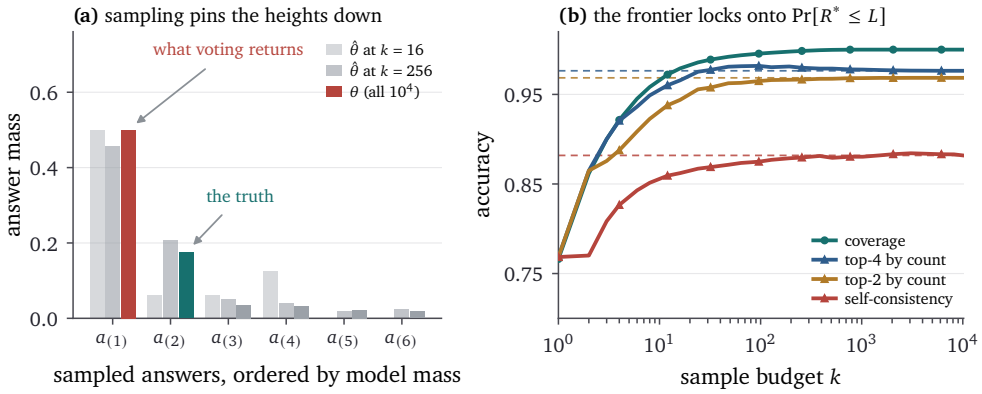


Figure 2: **What a sample set identifies. (a)** One GSM8K problem under Llama-3-8B. More samples pin the bar heights down; no height says which bar is true, and here the truth is the runner-up. **(b)** The consequence, measured to $k{=}10^4$: coverage keeps rising while each top-L count rule locks onto $\mathbb{P}[R^* \le L]$ (dashed). Selection plateaus not because voting is a poor rule but because there is nothing left to estimate. Only $L{=}1$ is an achievable selector accuracy; the $L{>}1$ curves are the value of a *count-ranked* shortlist, realizable with external information that resolves it.

1. **A rank law.** No label-free selector, at any budget, exceeds the identification ceiling set by the model's own per-problem output law, and for any feature map read per sample the ceiling tightens to that feature's own (Theorem 2.1). For answer counts we characterize the ceiling exactly: the entire count-ranked shortlist frontier is a functional of the truth's rank in the model's answer-mass ordering, while the exact oracle value of an M-message target-aware channel is the Bayes list ceiling G_M; we name the calibration condition under which plurality is the best count rule and the two agree (Corollary 2.4), with exponential finite-k rates (Theorem 2.6). This contains the modal ceiling of Bay & Yearick (2026) as the case $L{=}1$.
2. **A zero-free-parameter test.** Across 10 model–task cells, three model families and budgets spanning $k{=}1$ to 10^4, a θ fitted on half of a problem's samples predicts the other half's measured selection at every budget and depth with median absolute error 0.002 (Figure 3).
3. **The gap priced in oracle bits, and difficulty rather than benchmark identity organizes the price.** On the 12 difficulty strata where the gap is bridgeable, a count-ranked shortlist plus one ideal bit closes 0.36–1.00 of it and five bits 1.00; on the hard tail (per-problem pass@1 of 0.001–0.041) the truth's median rank on failures is 4–76 and one bit closes 0.00–0.15. The cell-level GSM8K/MATH separation does not survive matching on difficulty. What an implementable verifier collects of the price is far less, and negative once the verifier is the sampler.

Scope. Every bound is over label-free selectors reading the k samples of a *single* problem on checkable-answer tasks. External verifiers, tools and cross-problem constraints add information and exceed the ceiling, which is the point: Corollary 2.5 prices that escape rather than forbidding it.

2 The Rank Law

Setup. A problem is a pair (a^*, θ) drawn from a prior $\mathcal{D}$, where a^* lies in a finite answer alphabet A obtained by clustering equivalent outputs, and $\theta \in \Delta(\mathsf{A})$ is the model's per-problem output distribution. Nothing assumes θ puts its mass near a^*; the prior carries the whole relationship between what the model believes and what is true.

Assumption 1 (Conditional i.i.d. sampling)**.** Conditional on the problem, the k samples are i.i.d. draws from θ.

This is the only modelling assumption, and §4 tests it rather than asserting it. Note what it does *not* say: the correlation that defeats voting is not dependence between samples but the prior $\mathcal{D}$ placing

mass on problems whose θ concentrates on a shared mistake, correlated-jury failure (Ladha, 1992; Berg, 1993; Boland, 1989) in the form it takes here.

A *label-free selector* at budget k reading feature map Φ is any randomized map from $\Phi(\text{samples})$ to a sampled answer that never reads a^*; write $\mathcal{T}_k(\Phi)$ for the supremum of $\mathbb{P}[g = a^*]$ over such g. The range constraint is enforced throughout: a selector may return only an answer the model produced. Write $\mathsf{A}(\theta) = \{a : \theta(a) > 0\}$ for the answers the model can emit at all. Sorting A by θ descending, $\theta_{(1)} \geq \theta_{(2)} \geq \cdots$, define the *mass rank* $R^* = \min\{j : a_{(j)} = a^*\}$, with $R^* = \infty$ when $\theta(a^*) = 0$, so $\{R^* < \infty\} = \{a^* \in \mathsf{A}(\theta)\}$. Call Φ *per-sample* if it summarizes the sample set through the multiset of a fixed statistic of each individual sample, and *answer-preserving* if that statistic determines the sample's own answer; answer counts and any (answer, per-trace score) pair are both, and we use no other kind.

Theorem 2.1 (Identification ceiling)**.** *Let ν be the model's per-problem law of a single sample, let $\mathsf{A}(\nu) \subseteq \mathsf{A}$ be the answers it gives positive mass, and write $V^*(\nu) = \mathbb{E}[\max_{a \in \mathsf{A}(\nu)} \mathbb{P}[a^*{=}a \mid \nu]]$, the posterior ceiling under the* sampled-support convention*: the maximum ranges over answers the model can emit, not over all of A. Under Assumption 1, for every k, every label-free selector reading the sample set in any way attains at most $V^*(\nu)$. If in addition Φ is per-sample and answer-preserving, so that $\mathsf{A}(\nu)$ is a function of the induced law μ_Φ of one Φ-value, and $V^*(\Phi) = \mathbb{E}[\max_{a \in \mathsf{A}(\nu)} \mathbb{P}[a^*{=}a \mid \mu_\Phi]] \leq V^*(\nu)$, then*

$$\mathcal{T}_k(\Phi) \;\leq\; V^*(\Phi) \quad \textit{for every } k, \qquad \textit{and} \qquad \lim_{k \to \infty} \mathcal{T}_k(\Phi) \;=\; V^*(\Phi)$$

whenever μ_Φ is identifiable from its empirical measure and $a \mapsto \mathbb{P}[a^*{=}a \mid \mu_\Phi]$ is $\mathcal{D}$-a.s. continuous.

The proof (Appendix C) is short. Assumption 1 makes $a^* \to \nu \to$ samples a Markov chain, so the sample set is a noisy observation of ν and of nothing else about a^*; conditioning on ν then bounds any selector by the best answer *it is allowed to return*, and attainment follows because the empirical measure converges. Three scope remarks. The universal half needs no structure on what the selector computes, but it bounds only by $V^*(\nu)$; sharpening it to a *coarser* feature's own ceiling is what requires Φ to act per sample. Neither inequality needs regularity: the two displayed conditions buy attainment only, and for answer counts they are immediate, θ on a finite alphabet being identified by its empirical measure. That is the feature map every measurement below uses.

Third, the sampled-support convention is load-bearing, not cosmetic: a selector returns a sampled answer while the unrestricted Bayes rule need not, so maximizing over all of A would break the limit and would push the ceiling above coverage in Corollary 2.2 (Appendix C gives the example; the two conventions coincide under mass calibration and differ by at most $1 - \mathrm{Cov}_k$). The content is in what the theorem forbids: neither ceiling depends on k, so no selector, however clever, escapes by sampling more.

Corollary 2.2 (The measurable frontier)**.** *Take Φ to be answer counts and write $S_L(k) = \mathbb{P}[a^* \in \text{top-}L \text{ by count}]$. If θ is a.s. untied at positions $L, L{+}1$ then $S_L(k) \to \mathbb{P}[R^* \leq L]$, and*

$$\underbrace{\mathbb{P}[R^*{=}1]}_{\textit{self-consistency}} \;\leq\; \underbrace{V^*(\textit{counts})}_{\textit{best count rule}} \;\leq\; \underbrace{\mathbb{P}[R^*{<}\infty]}_{\textit{coverage}}.$$

The left-hand quantity is the modal ceiling of Bay & Yearick (2026). The two inequalities are this paper's two empirical questions: how far above the modal ceiling V^* sits (§6: measurably, not at all), and how fast $\mathbb{P}[R^* \leq L]$ climbs (§5).

The left inequality in Corollary 2.2 is where every label-free selector paper lives, and an unbounded search for a better rule can only ever lower-bound V^*. We close it instead by naming the condition. Fix a budget k and let Φ be any observable summary of the counts. A count rule acts by naming a *position* in the observed count ordering, so write

$$\pi_j(\Phi) \;=\; \mathbb{P}[\text{the truth occupies position } j \text{ of the observed count ordering} \mid \Phi],$$

and call the model *rank-calibrated at* Φ if $\pi_j(\Phi)$ is non-increasing in j almost surely. (A finite-k statement about the *observed* ordering, which is what a selector acts on; in the limit the two orderings coincide.)

Proposition 2.3 (Rank calibration). *At any budget k, $\mathcal{T}_k(\Phi) - \mathbb{P}[\textit{plurality}{=}a^*] = \mathbb{E}_\Phi[\max_j \pi_j(\Phi) - \pi_1(\Phi)]$, so answering the plurality attains $\mathcal{T}_k(\Phi)$ if and only if $\pi_1(\Phi) = \max_j \pi_j(\Phi)$ almost surely. Rank calibration implies that, and delivers the same conclusion at every shortlist depth (Corollary 2.4); the converse fails, since only position 1 is constrained here. If the model is mass-calibrated, $\mathbb{P}[a^*{=}a \mid \theta] = \theta(a)$, then $\mathbb{P}[a^*{=}a_{(j)} \mid \theta] = \theta_{(j)}$ is non-increasing in j, so answering the highest-mass answer is Bayes-optimal given θ and, by Corollary 2.2, the plurality attains $V^*(\textit{counts})$ in the limit.*

The proposition is elementary and that is its use: any label-free rule that beats the plurality is a detector of one specific direction of miscalibration, so §6 can *estimate* π_j rather than run a bake-off. The same argument runs for shortlists, and that is what ties the count-ranked frontier to the Bayes value of external information.

Corollary 2.4 (List rank calibration). *Write $G_L(\Phi)$ for the* Bayes list ceiling*, the largest probability that a Φ-measurable shortlist of at most L sampled candidates contains a^* (Theorem B.1); $G_1(\Phi) = \mathcal{T}_k(\Phi)$ and $G_L(\Phi) \geq S_L(k)$ always. If Φ determines the observed ordering and the model is rank-calibrated at Φ, then for every L the Bayes L-element shortlist is that ordering's first L positions, so $G_L(\Phi) = S_L(k)$; for a fixed L it is enough that those L positions are a.s. the L largest $\pi_j(\Phi)$.*

Corollary 2.5 (The gap has a price in bits, as oracle capacity). (i) *An external* target-aware, noiseless *channel whose per-problem transcript takes at most M values raises the attainable value to exactly $G_M(\Phi)$, the expected top-M posterior mass given Φ: a capacity, attained only by a channel that already knows the truth, with $M{=}2^B$ the exact value of B bits.* (ii) *One such channel shortlists the top M answers by count and names the correct member whenever the truth lies there, attaining $S_M(k) \leq G_M(\Phi_{\text{counts}})$, hence $\mathbb{P}[R^* \leq M]$ in the limit.* (iii) *Under Corollary 2.4 the two coincide, so that realization is Bayes-optimal. The gap's* count-ranked width *at tolerance ε is $M^*_{\text{rank}} = \min\{L : S_L(k) \geq \text{Cov}_k - \varepsilon\}$, with limiting form $\min\{L : \mathbb{P}[R^* \leq L] \geq \mathbb{P}[R^*{<}\infty] - \varepsilon\}$, in bits $B^*_{\text{rank}} = \log_2 M^*_{\text{rank}}$; the* Bayes message width*, the same minimum over G_L, is never larger, equal under* (iii).

This is the theory's diagnostic content: it converts *"the gap is large"* into a width in bits, and §5 shows those are very different statements. Part *(i)* binds any bounded-message target-aware channel exactly, part *(ii)* is the width §5 measures, and neither promises that a deployable mechanism collects the bits: they price an idealized channel, not model calls, and §7 measures what a real verifier collects. The approach to every limit is exponential, which makes a k-sweep a quantitative rather than qualitative prediction.

Theorem 2.6 (Finite-k rate). *Fix a problem with r wrong answers of positive mass and let $m_L = \theta(a^*) - \theta_{(L)}$ be the margin against the L-th largest wrong mass, with $m_L \neq 0$. Then $\left|\mathbb{P}[a^* \in \text{top-}L] - \mathbf{1}\{m_L > 0\}\right| = e^{-k\Lambda_L(1+o(1))} \leq \max(r, L)\, e^{-km_L^2/2}$, where Λ_L is the exact Sanov rate of the relevant event (I_L of* (3) *when $m_L > 0$, its mirror $\tilde{I}_L$ when $m_L < 0$).*

Where $m_L < 0$, additional samples do not merely fail to help: they make the wrong answer win with exponentially increasing certainty.

3 Measurement

Cells. Ten model–task pairs, called *cells* throughout, in two complementary families. The *pool* cells reuse the *Large Language Monkeys* release (Brown et al., 2024) (Llama-3-8B/70B-Instruct and Gemma-2B/7B on GSM8K and MATH, $k{=}10^4$ samples on $n{\approx}128$ problems) and carry the budget axis and cross-family breadth. The *grid* cells are our own Qwen2.5-3B/7B-Instruct (Qwen Team, 2024) runs on *full* GSM8K ($n{=}1319$) and MATH-500 ($n{=}500$) at $k{=}64$, $T{=}0.8$, seed 0, and carry benchmark-scale problem counts. Neither family alone is sufficient and we do not pool them.

Count banks, and the label-free discipline. Every cell is reduced to one object: per problem, a cluster id per sample plus the id of the correct cluster. To prevent target leakage, clusters are built from sampled strings alone by a canonicaliser that never sees a gold answer, and gold correctness is attached only after the partition is fixed. Under-measuring coverage would shrink the gap's denominator and inflate the reported bit price, so the truth cluster is named by the *union* of two independent oracles, our `math_verify` re-grade and the release's per-sample labels. Appendix D gives the clustering and the gates cell by cell.

Table 1: **The frontier and its limits, on every cell.** $\mathbb{P}[R^* \leq L]$ is estimated on the full bank; brackets are bootstrap 95% CIs over problems. "gap" is coverage $-$ $\mathbb{P}[R^*{=}1]$, and "1-bit" the share of it closed by moving from $L{=}1$ to $L{=}2$. Pool cells have large k and $n \approx 128$; grid cells have full benchmark n and $k{=}64$.

					$\Pr[R^* \leq L]$					
task	model	n	k	pass@1	$L{=}1$	$L{=}2$	$L{=}4$	coverage	gap	1-bit
GSM8K	Llama-3-8B	127	10000	0.766	0.882 [0.819, 0.937]	0.969	0.976	1.000	0.118	0.73
GSM8K	Llama-3-70B	127	10000	0.930	0.969 [0.937, 0.992]	0.992	1.000	1.000	0.031	0.75
MATH	Llama-3-8B	128	10000	0.281	0.445 [0.359, 0.531]	0.531	0.656	0.984	0.539	0.16
MATH	Llama-3-70B	128	10000	0.494	0.609 [0.523, 0.695]	0.758	0.836	0.984	0.375	0.40
MATH	Gemma-7B	128	10000	0.216	0.422 [0.336, 0.508]	0.531	0.648	0.984	0.562	0.19
MATH	Gemma-2B	128	10000	0.103	0.227 [0.156, 0.297]	0.320	0.406	0.953	0.727	0.13
GSM8K	Qwen2.5-3B	1319	64	0.827	0.908 [0.892, 0.923]	0.954	0.967	0.985	0.077	0.60
GSM8K	Qwen2.5-7B	1319	64	0.891	0.925 [0.911, 0.939]	0.967	0.976	0.983	0.058	0.74
MATH-500	Qwen2.5-3B	500	64	0.601	0.724 [0.682, 0.762]	0.794	0.860	0.916	0.192	0.36
MATH-500	Qwen2.5-7B	500	64	0.672	0.768 [0.732, 0.804]	0.850	0.884	0.936	0.168	0.49

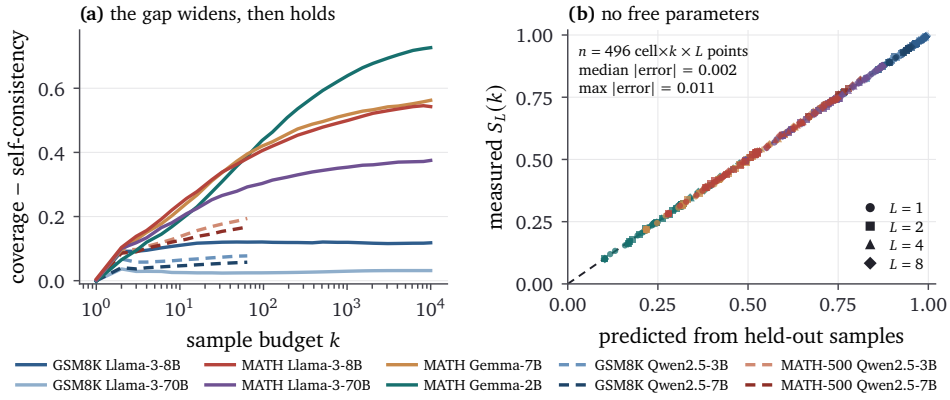


Figure 3: **The rank law, measured and predicted. (a)** The coverage–selection gap against budget on every cell: it widens as coverage climbs, then holds at the value Corollary 2.2 predicts. Solid lines are pool cells (k to 10^4), dashed the grid cells. **(b)** Every cell $\times$ budget $\times$ depth point, predicted against measured, fitting θ on half of each problem's samples and evaluating on the disjoint half, with no fitted parameter.

Statistics. Bootstrap 95% CIs over problems, ≥ 2000 resamples. Budgets below a bank's size are obtained by subsampling it without replacement, 128 subsamples per problem per budget. Selectors never read gold labels; gold labels only measure.

4 The frontier is the rank distribution, and it is predicted

The limits are attained. Figure 2(b) shows the mechanism on one cell and Table 1 gives it on all ten. On GSM8K/Llama-3-8B, self-consistency settles at 0.882 against a predicted $\mathbb{P}[R^*{=}1]$ of 0.882 and top-2 at 0.969 against a predicted 0.969, while coverage reaches 1.000 by $k = 10^4$: all 127 problems are solved by at least one sample, so every remaining self-consistency error there is a selection failure and not a capability failure.

The prediction is quantitative and out of sample. Estimating θ from the samples that also decide the vote is circular, so we split each problem's bank in half, fit θ on one half, and predict the measured selection curve on the disjoint other half at every budget k and depth L, with nothing fitted. Across 496 such points (Figure 3b) the median absolute error is 0.002 and the maximum is 0.011: Assumption 1 and the rank law reproduce the answer-count frontier itself, not only its shape.

The law survives intervention. Holding checkpoint, prompts and decoding fixed, enabling Qwen3-8B thinking raises pass@1 by 6.35 points and $\mathbb{P}[R^*{\leq}1]$ by 5.18; the independently measured selection-frontier shift matches the held-out rank prediction under its calibrated null (Appendix P.10).

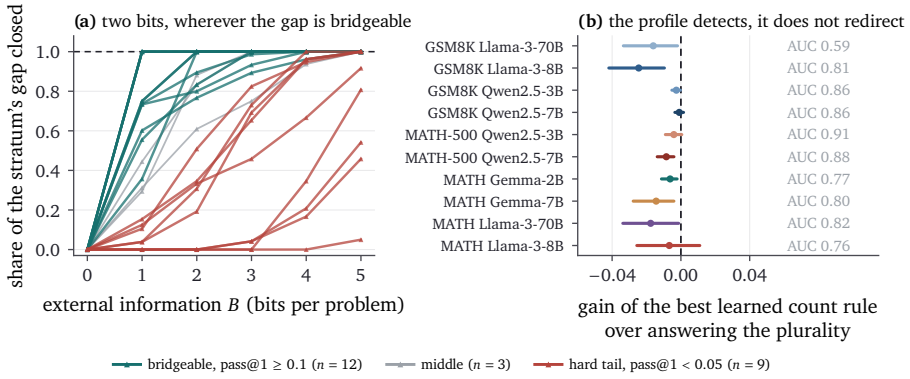


Figure 4: **The price of the gap is organized by difficulty, not by benchmark identity. (a)** Share of each stratum's coverage–selection gap closed by B bits from an idealized target-aware oracle spent inside the count-ranked top-2^B shortlist, $\mathbb{P}[R^* \leq 2^B]$ normalised between self-consistency and coverage, one curve per difficulty stratum: the rank realization of Corollary 2.5*(ii)*. These are capacities, not verifier accuracies (§7). Where the gap is bridgeable at all it closes within two to five bits whatever the benchmark, model family or scale; on the hard tail it barely moves. **(b)** A cross-fitted gradient-boosted rule over count-profile features, against simply answering the plurality. It detects when the plurality is wrong (AUC at right) and never profitably answers anything else.

5 The price of the gap: Bayes capacity and its rank realization

Corollary 2.5 gives the gap a width in bits; Figure 4(a) measures it, in oracle capacity throughout: a small price means the gap is cheap *in principle*, not closed. Every number below estimates $\mathbb{P}[R^* \leq 2^B]$, so it is the count-ranked width B^*_{rank}, which bounds the Bayes width above. Read at cell level the picture looks like a task property: one bit closes 0.60–0.75 of the gap on the GSM8K cells and 0.13–0.49 on the MATH cells, with no overlap. That reading does not survive stratification, and correcting it is the more useful result.

The cell-level separation goes with difficulty composition, not with the benchmark. We split every cell into five equal-count bins of per-problem pass@1 and recompute the one-bit share within each (Table 3). First, within each of the six cells contributing more than one graded stratum, the share is non-decreasing in the bin's pass@1: difficulty orders the price inside a cell. Second, the benchmark ordering does not survive matching. Over the range of estimated pass@1 where the two task families' bridgeable strata overlap (0.26 to 0.50), the MATH strata run 0.75 to 1.00 against 0.60 to 0.74 for the GSM8K strata: not systematically below, and mostly above, so what differs is the difficulty mix and not the price. Call a stratum *bridgeable* when its pass@1 exceeds 0.10. Across the 12 bridgeable strata (both task families, three model families, five model sizes), a count-ranked shortlist plus one ideal bit closes 0.36–1.00 of the gap, two bits 0.77–1.00, and five bits 1.00 of it. Difficulty here is estimated from the samples that also define R^*; cross-fitting removes that circularity, cutting the bins on one half of each problem's samples and measuring the price on the other. Every statement in this paragraph survives: the price stays monotone in difficulty on 6 of 6 cells, the matched bands still overlap, and no stratum's share moves by more than 0.13 (Appendix E.1).

Where voting fails, the truth is the runner-up. The price is small for one reason: on 92% of those strata the median rank of the truth *among problems where the plurality is wrong* is exactly 2, and across all of them it lies in 2.0–3.0. When self-consistency fails on a problem the model can solve, it usually fails by one position.

The exception, and it is the important one. The share falls off continuously as strata get harder, and at the bottom it is near zero: on the nine strata with per-problem pass@1 of 0.001–0.041, the truth's median rank on failures is 4–76 and one bit closes only 0.00–0.15 of the gap. There the truth is not the runner-up; it is buried, or absent from the sample set entirely. This is the same difficulty dependence Snell et al. (2024) found for allocation, now on the selection side and denominated in bits.

Is this just accuracy? Partly, and that is the point: the bit price tracks per-problem difficulty rather than the benchmark label, so it must be measured per stratum. The cell-level numbers in Table 1 are what a practitioner running that exact cell would face, but they do not transfer to a different difficulty mix (Appendix E). The mechanism is the error structure: where a model's errors collapse onto a single attractor the truth survives in a two-element count shortlist and one bit suffices, and where they are multi-modal the truth escapes it. On our grid cells the mean number of distinct sampled answers is 4.1 and 2.3 on GSM8K against 10.0 and 6.8 on MATH-500, the same fact in another coordinate.

6 Why a better label-free rule is not the answer

By Proposition 2.3 there is exactly one thing to measure: the posterior $\pi_j(\Phi)$ over which position holds the truth, and whether position 1 is its maximum. We estimate it with a class far richer than the count rules the literature proposes. From the count profile alone (top-1/2/3 shares, their ratios, normalised entropy, Simpson index, singleton fraction, support size, budget), a gradient-boosted classifier is cross-fitted over problems and the induced selector answers its arg-max position. Its held-out accuracy is an attainable *lower* bound on $V^*(\text{counts})$, never an upper bound; the fitted π_j supplies in addition the calibration-violation estimate that Proposition 2.3 prices.

The result splits cleanly in two (Figure 4b). *Detection works*: the profile predicts whether the plurality is correct with AUC 0.59–0.91, consistent with the confidence literature (Ding, 2026). *Switching never pays*: on every one of the ten cells the learned rule is *worse* than answering the plurality, with per-cell differences of -0.025–-0.001, the interval excluding zero on 7 of them and pooled at -0.006 $[-0.008, -0.003]$. The fitted posterior does favour position 2 on 2–23% of profiles, but acting on those is wrong 64–92% of the time, the signature of estimation noise rather than of recoverable signal. Seven pre-registered trace-aware selectors, including surprisingly-popular (Prelec et al., 2017) and confidence-weighted voting (Tian et al., 2023), reproduce this on the four grid cells, where the tests had 80% power against improvements of $+0.16$ to $+0.19$ over the count baseline and none rejected (Appendix P).

By Proposition 2.3 any rule that beat the plurality would have had to detect a profile on which position 1 is not the posterior-most-likely one, and none we tested does at these sample sizes. That is sharper than a bake-off null: it measures the quantity the proposition says is decisive, across the count-profile family in full and a pre-registered trace-feature family, and it comes out at zero. Within those families, sampling tells you *how confident to be* and not *what else to say*, which is why the gap's price is denominated in external bits.

7 What an implementable verifier collects

The bit price is an information bound attained by a target-aware oracle; what a deployable verifier realizes of it is a separate, measurable question, and the distance between the two is the practical issue. Every verifier below sees only the problem and a count-ranked shortlist of L candidates in deterministically randomized order, optionally with one representative trace each, and must return an index; a model call is never called "a bit". All conditions run at $k{=}16$ on the multi-candidate slice with problem-level paired bootstrap (Appendix O).

The oracle headroom is real: on the GSM8K/Qwen2.5-3B slice the plurality is right 0.803 of the time while the truth is in the count-ranked top-2 shortlist 0.905 of the time, so a perfect one-of-two selector would gain 10.2 points. Of the 26 verifier conditions we ran (four verifier roles, two depths, two prompt forms, four cells), exactly one collects any of it (Figure 5). *Index-only* verifiers do not merely fail: on GSM8K/Qwen2.5-3B they lose 22 to 50 points against a count prior already right four times in five, picking position 1 between 16% and 81% of the time under randomized order. Letting the same 7B model reason before committing reverses the sign once, to $+0.031$ $[+0.005, +0.057]$ on GSM8K/3B, about 30% of the available top-2 headroom, at a position-1 rate of 0.49. Two controls dissolve even that. Matched on measured tokens, sampling the 7B model eight times reaches 0.918 for $1{,}154$ tokens while 3B-sample-then-7B-verify reaches 0.913 for $3{,}120$: the hybrid is Pareto-dominated by using the better generator. And with the verifier equal to the sampler (the case that matters, one already running one's best model), it falls 10.9 points below plurality on GSM8K and 15.3 on MATH-500, while reasoning and showing no position bias.

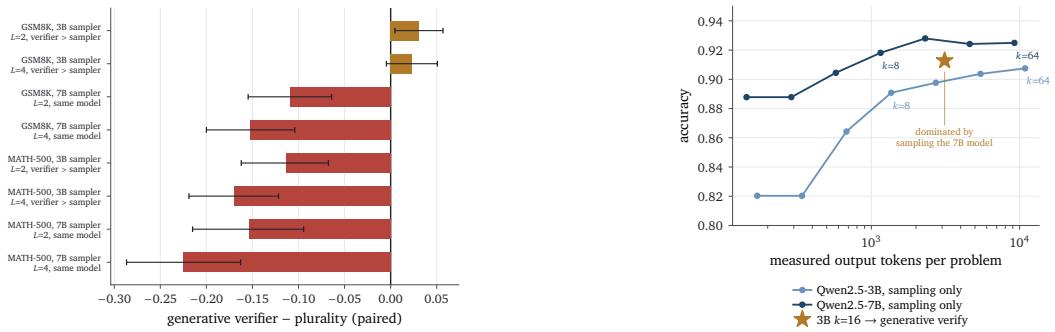


Figure 5: **What a real verifier collects of the bit price.** *Left:* the reasoning Qwen2.5-7B verifier, k=16, paired bootstrap. Only the condition whose verifier outranks its sampler beats plurality; L=4 loses to L=2 throughout. *Right:* accuracy against output tokens. Verification beats more 3B sampling but is dominated by sampling the 7B generator directly.

A same-model verification prompt need not induce the same θ; empirically, it overrides a k-draw count prior with a single-pass judgement. Oracle capacity is substantial and one condition realizes about a third, but verifier–sampler capability asymmetry explains that gain: generation closes it more cheaply, as cross-model and self-scoring results also find (Singhi et al., 2025; Liu, 2026). Appendix O reports every condition and the pre-registered secondary tests.

8 Related work

The phenomenon. Brown et al. (2024) established the coverage law and plateau; Bay & Yearick (2026) formalise its L=1 ceiling. We give the full rank CDF, finite-k law and bit price. Shared misconceptions limit aggregation (Denisov-Blanch et al., 2026), agreement-based confidence (Ding, 2026), and effective judge count (Kohli, 2026).

Selection and verification. Theorem 2.1 jointly bounds plurality (Wang et al., 2023), surprisingly-popular (Prelec et al., 2017), weighted voting (Tian et al., 2023; Nitzan & Paroush, 1982), step-aware reranking (Li et al., 2023), and set-level Universal Self-Consistency (Chen et al., 2023); label-supervised reward models leave that class (Lightman et al., 2024). Adaptive allocation is prior art (Singhi et al., 2025; Zhai et al., 2026; Dughmi et al., 2026; Wang et al., 2025); its difficulty dependence (Snell et al., 2024) appears here in bits.

Classical machinery. We instantiate correlated-jury failure (Ladha, 1992; Berg, 1993; Boland, 1989), guessing probability and min-entropy (Cover & Thomas, 2006; Dodis et al., 2008), Blackwell comparison (Blackwell, 1953) and list decoding in test-time sampling.

9 Discussion

What to measure and build. Sample a small labelled probe at generous k, histogram the truth's count rank, and read off the count-ranked width $B^*_{\mathrm{rank}}(\varepsilon)$ by difficulty. Where it is near one, expose a two-element shortlist for an external check; where it is large, use a better or differently-correlated generator. In our one verifier gain, the stronger generator paid more.

Limitations. We test checkable-answer math, mostly instruction-tuned models, with full-benchmark runs ending at k=64, one Qwen3-8B thinking intervention, and verifiers up to 7B in one family. Prices need remeasurement for other reasoning models and open-ended tasks; negatives bound only the tested families (Appendix N).

10 Conclusion

Repeated sampling learns what a model believes, not what is true. The truth's rank governs the coverage–selection gap; escaping it costs external information set by difficulty, a price our verifiers largely fail to pay.

Reproducibility Statement

The sampling protocol (models, datasets, budgets, temperature, seeds, token caps and evaluation rules), the answer-clustering and correctness procedures, the statistical analyses, and the scope conditions are described in the main text and appendices; the pre-registered predictions and every subsequent deviation from them are reported in Appendix P. Proofs of the theoretical results, together with the independent numerical checks of those results and the remaining experimental details, are given in the appendices.

AI Use Statement

AI-assisted language tools were used for editing and improving the readability of the manuscript. They were not used for research ideation, methodology, mathematical derivations, experimental design, or interpretation of results. The authors reviewed all AI-assisted edits and take responsibility for the final manuscript.

References

Erdal Arıkan. An inequality on guessing and its application to sequential decoding. *IEEE Transactions on Information Theory*, 42(1):99–105, 1996.

Erdal Arıkan and Neri Merhav. Guessing subject to distortion. *IEEE Transactions on Information Theory*, 44(3):1041–1056, 1998.

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.

Sven Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1): 87–95, 1993.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Philip J. Boland. Majority systems and the Condorcet jury theorem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 38(3):181–189, 1989.

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.

Antonio Cabrales, Olivier Gossner, and Roberto Serrano. Entropy and the value of information for investors. *American Economic Review*, 103(1):360–377, 2013.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Yegor Denisov-Blanch et al. Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness. *arXiv preprint arXiv:2603.06612*, 2026.

Kaihua Ding. When LLMs agree, are they right? Auditing self-consistency and cross-model agreement as confidence signals. *arXiv preprint arXiv:2607.08065*, 2026.

Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM Journal on Computing*, 38(1): 97–139, 2008.

Shaddin Dughmi, Mahdi Haghifam, and Yusuf Hakan Kalayci. Adaptive generate-rank-verify: Inference-time search with costly verification. *arXiv preprint arXiv:2605.17609*, 2026.

G. David Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, 1968.

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*, 2021. arXiv:2103.03874.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026.

Krishna K. Ladha. The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634, 1992.

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. *Proceedings of ACL*, 2023. arXiv:2206.02336 (DiVeRSe).

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 (ICLR)*, 2024. arXiv:2305.20050.

Ning Liu. LLMs as a jury: Cross-model consensus can outperform process reward models for LLM reasoning. *arXiv preprint arXiv:2607.10139*, 2026.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

James L. Massey. Guessing and entropy. In *Proceedings of the IEEE International Symposium on Information Theory*, pp. 204, 1994.

Giuseppe Moscarini and Lones Smith. The law of large demand for information. *Econometrica*, 70 (6):2351–2366, 2002.

Xiaosheng Mu, Luciano Pomatto, Philipp Strack, and Omer Tamuz. From Blackwell dominance in large samples to Rényi divergences and back again. *Econometrica*, 89(1):475–506, 2021.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982.

Dražen Prelec, H. Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

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

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

Seymour Sherman. On a theorem of Hardy, Littlewood, Polya, and Blackwell. *Proceedings of the National Academy of Sciences*, 37(12):826–831, 1951.

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. *arXiv preprint arXiv:2504.01005*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of EMNLP*, 2023. arXiv:2305.14975.

Erik Torgersen. *Comparison of Statistical Experiments*, volume 36 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1991.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. DynScaling: Efficient verifier-free inference scaling via dynamic and integrated sampling. *arXiv preprint arXiv:2506.16043*, 2025.

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 (ICLR)*, 2023. arXiv:2203.11171.

Nir Weinberger and Ofer Shayevitz. Guessing with a bit of help. *Entropy*, 22(1):39, 2020.

Zhiyuan Zhai, Bingcong Li, Bingnan Xiao, Ming Li, and Xin Wang. Adaptive test-time compute allocation for reasoning LLMs via constrained policy optimization. *arXiv preprint arXiv:2604.14853*, 2026.

Appendices A–C develop the notation and the core proofs; D–E cover measurement and difficulty; F–L provide the supporting theory and analyses; M–N discuss reproducibility and scope; O reports the verifier experiments; and P contains the extended experiments and the thinking-mode intervention.

A Notation: the rank and guessing-curve vocabularies

The body states everything in terms of the mass rank R^*; the proofs and the extended report are written in the guessing-curve vocabulary, which is the one the classical results are stated in. The two name the same objects, and this table is the dictionary. Throughout, Φ is a per-sample feature map, k the budget, and every quantity is an average over the problem prior $\mathcal{D}$.

Object	Guessing-curve name	Rank name
plurality accuracy at budget k	F_k, frequency ceiling	$\to \mathbb{P}[R^*{=}1]$ as $k \to \infty$
top-L-by-count success at budget k	$S_L(k)$, list envelope	$\to \mathbb{P}[R^* \le L]$
best count rule at budget k	$T_k(\Phi_{\mathrm{ct}})$	$\to V^*(\text{counts})$, the count ceiling
best rule on any per-sample feature	$T_k(\Phi)$	$\to V^*(\Phi)$
coverage at budget k	Cov_k	$\to \mathbb{P}[R^* {<} \infty]$
Bayes value of a label-free L-shortlist	$G_L(\Phi)$, guessing curve	$\ge S_L(k)$; equal under Cor. 2.4
exact oracle value of $M{=}2^B$ messages	$G_M(\Phi)$	$\ge S_M(k) \to \mathbb{P}[R^* \le M]$

Two distinctions are load-bearing and are never elided. *Finite-k versus limit:* R^* is the rank by *mass*, a property of the problem; what a selector observes is the rank by *count* at budget k, and the two coincide only in the limit (Theorem B.2 gives the exponential rate). Tables reporting "rank 2" at a stated k report the observed count rank. *Bayes versus count-ranked:* $G_{2^B}(\Phi)$ is *exactly* what a target-aware channel of 2^B transcripts could buy, whereas $S_{2^B}(k) \to \mathbb{P}[R^* \le 2^B]$ is what one particular realization of that channel buys, the one that spends its bits inside the count-ranked top-2^B shortlist; the second never exceeds the first and equals it under Corollary 2.4. Every empirical bit price in this paper is the count-ranked quantity. *Oracle versus implementable:* both are capacities; what a model call collects is measured separately in Appendix O and is much less. A model call is never called "a bit".

B The guessing-curve formulation

The rank formulation in the main text admits an equivalent representation through the Bayes *guessing curve* $G_L(\Phi)$, the expected posterior mass of the L most probable sampled answers. We state that formulation here for reference; several appendices below prove or use it.

Theorem B.1 (The guessing curve)**.** *For every feature map Φ of the sample set:* ***(i)*** $G_1(\Phi) = \mathcal{T}(\Phi) = \mathbb{E}_\Phi[\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^*{=}a \mid \Phi]] \le 2^{-H_\infty(a^* \mid \Phi)}$*, the Arimoto conditional min-entropy bound, with difference at most* $1 - \mathrm{Cov}_k$*. Here and throughout,* $\mathsf{A}(\Phi)$ *is the set of sampled answers and*

every maximum and every shortlist is confined to it: this is the sampled-support convention of Theorem 2.1, forced by the requirement that a selector return an answer the model produced. ***(ii)*** *$G_L(\Phi)$ is the maximal probability that a label-free L-element shortlist built from Φ contains a^*, and $G_L \uparrow \mathrm{Cov}_k$ once L exhausts the sampled candidates.* ***(iii)*** *For every $M \geq 1$, the supremum of $\mathbb{P}[\text{selection} = a^*]$ over target-aware oracle channels whose per-problem transcript V satisfies $|\mathrm{range}(V)| \leq M$ is exactly $G_M(\Phi)$; specializing to $M = 2^B$ gives the value of B bits, which is Corollary 2.5*(i)*. The count-ranked realization of Corollary 2.5*(ii)* attains $S_M(k) \leq G_M$, with equality under Corollary 2.4.* ***(iv)*** *If $a^* \to \Phi \to \Psi$ is Markov then $G_L(\Psi) \leq G_L(\Phi)$ for every L.* ***(v)*** *The increments are nonincreasing, so $G_{2^B} \leq \min(1, 2^B G_1)$.*

Part (iv) of Theorem B.1, applied to the coarsenings traces $\to$ counts $\to$ plurality, gives the chain

$$\mathrm{Cov}_k \;\geq\; T(\Phi_{\mathrm{traces}}) \;\geq\; T(\Phi_{\mathrm{counts}}) \;\geq\; F, \tag{1}$$

which splits the coverage–selection gap into a level available to a better count-only rule and a trace level above it. In the rank presentation of the body the same chain reads $\mathbb{P}[R^*{<}\infty] \geq V^*(\text{traces}) \geq V^*(\text{counts}) \geq \mathbb{P}[R^*{=}1]$.

Theorem B.2 (Rank zero–one law)**.** *Fix a problem with r wrong answers of positive mass, under Assumption 1. For every L: if $m_L > 0$ then $1 - \mathbb{P}[a^* \in \text{top-}L \text{ counts}] = e^{-kI_L(1+o(1))} \leq r\, e^{-km_L^2/2}$; if $m_L < 0$ then $\mathbb{P}[a^* \in \text{top-}L \text{ counts}] = e^{-k\tilde{I}_L(1+o(1))} \leq L\, e^{-km_L^2/2}$, with the exact Sanov rates* (3)*. Hence $\mathbb{P}[a^* \in \text{top-}L \text{ counts}] \to \mathbf{1}\{L \geq R^*\}$. This is Theorem 2.6 of the body together with Corollary 2.2.*

Theorem B.3 (Parameter-free margin law)**.** *Under Assumption 1 in the two-mode reduction, with $\sigma^2 = p + (1-p)w - m^2$, $F_k = \Phi(\sqrt{k}\, m/\sigma) + o(1)$, so per-problem voting accuracy against $z = \sqrt{k}\, m/\sigma$ collapses onto one universal curve with no free parameters.*

C Proofs for the rank law

C.1 Theorem 2.1

The universal bound. Let ν be the per-problem law of a single sample and write $p_\nu(a) = \mathbb{P}[a^*{=}a \mid \nu]$ and $\mathsf{A}(\nu) = \{a : \nu \text{ gives answer } a \text{ positive mass}\}$. By Assumption 1 the generative order is $(a^*, \nu) \sim \mathcal{D}$ and $s_1, \ldots, s_k \mid (a^*, \nu) \overset{\text{iid}}{\sim} \nu$, so the samples depend on a^* only through ν: the chain $a^* \to \nu \to (s_1, \ldots, s_k)$ is Markov, and it remains Markov after any measurable map of the sample set. Equivalently, $a^* \perp s_{1:k} \mid \nu$.

Let g be any label-free selector, computed from the samples in any way whatsoever, possibly randomized with randomness independent of $(a^*, \nu, s_{1:k})$. By definition g returns a *sampled* answer, so $g \in \mathsf{A}_k \subseteq \mathsf{A}(\nu)$ almost surely, where A_k is the set of answers actually observed. Conditioning on ν and using $a^* \perp s_{1:k} \mid \nu$,

$$\mathbb{P}[g = a^*] = \mathbb{E}\big[\mathbb{P}[a^*{=}g \mid \nu,\, s_{1:k}]\big] = \mathbb{E}\big[p_\nu(g)\big] \;\leq\; \mathbb{E}\big[\max_{a \in \mathsf{A}(\nu)} p_\nu(a)\big] = V^*(\nu),$$

the inequality being pointwise, since g takes values in $\mathsf{A}(\nu)$. This holds at *every* k and needs no regularity: it is the ceiling no amount of sampling escapes.

The restriction of the maximum to $\mathsf{A}(\nu)$ is not cosmetic. Dropping it gives a valid but strictly weaker bound (a maximum over A dominates one over $\mathsf{A}(\nu)$), and the weaker version is *not* attained: a selector confined to sampled answers cannot chase a posterior maximizer the model never emits. Take $\mathsf{A} = \{A, B\}$ with ν supported on B alone and $p_\nu(A) = 0.9$, $p_\nu(B) = 0.1$; the unrestricted maximum is 0.9 while $\mathbb{P}[g = a^*] = 0.1$ for every g and every k. With the convention in force, $V^*(\nu) = 0.1$ and the theorem's limit is correct. The prior is free to place mass on such problems, which is precisely the event $R^* = \infty$, so the convention, rather than an assumption ruling those problems out, is what keeps the statement both true and non-vacuous.

Sharpening to a coarser feature. Now let Φ act per sample and be answer-preserving, and let μ_Φ be the push-forward of ν, so the k observed feature values are i.i.d. μ_Φ. Answer-preservation makes the answer marginal of μ_Φ equal to that of ν, so $\mathsf{A}(\nu)$ is a function of μ_Φ and the index set of the restricted maximum is μ_Φ-measurable, which is what lets the argument run at the level of μ_Φ at

all. Conditional on μ_Φ, the law of $\Phi_{1:k}$ is $\mu_\Phi^{\otimes k}$ whatever a^* is, so $a^* \to \mu_\Phi \to \Phi_{1:k}$ is Markov as well, a selector reading $\Phi_{1:k}$ still returns an element of $\mathsf{A}(\nu)$, and the same conditioning argument gives $\mathcal{T}_k(\Phi) \le V^*(\Phi)$ for every k. For $V^*(\Phi) \le V^*(\nu)$, apply the tower rule and then Jensen for the convex map $x \mapsto \max_{a\in\mathsf{A}(\nu)} x_a$ on the fixed index set $\mathsf{A}(\nu)$: $\max_{a\in\mathsf{A}(\nu)} \mathbb{E}[p_\nu(a) \mid \mu_\Phi] \le \mathbb{E}[\max_{a\in\mathsf{A}(\nu)} p_\nu(a) \mid \mu_\Phi]$. A coarser reading of each sample can only lower the ceiling.

Attainment. The decoder must respect the range constraint, so take the plug-in rule restricted to what has been seen,

$$g_k = \arg\max_{a\in\mathsf{A}_k} \mathbb{P}[a^*{=}a \mid \hat{\mu}_k],$$

with A_k the observed answer set and $\hat{\mu}_k$ the empirical measure of Φ-values. Two convergences are needed and both hold. First, $\mathsf{A}_k \uparrow \mathsf{A}(\nu)$ a.s.: A is finite, each answer in $\mathsf{A}(\nu)$ has mass bounded below by some $\delta > 0$ under ν, and Borel–Cantelli makes the event that it is never drawn null. So the feasible set of g_k equals the index set of $V^*(\Phi)$ eventually, almost surely; this is the step the unrestricted decoder cannot supply, since its argmax may sit permanently outside A_k. Second, $\hat{\mu}_k \to \mu_\Phi$ weakly a.s. (Varadarajan), and by assumption μ_Φ is identifiable from $\hat{\mu}_k$ and $a \mapsto \mathbb{P}[a^*{=}a \mid \mu_\Phi]$ is $\mathcal{D}$-a.s. continuous in μ_Φ. Hence g_k agrees eventually a.s. with the support-restricted Bayes decoder $\arg\max_{a\in\mathsf{A}(\nu)} \mathbb{P}[a^*{=}a \mid \mu_\Phi]$, and bounded convergence gives $\mathbb{P}[g_k = a^*] \to V^*(\Phi)$. Together with $\mathcal{T}_k(\Phi) \le V^*(\Phi)$ this is the claimed limit. □

Exactly what is and is not claimed. The per-sample hypothesis is used only for the *sharpening*, and it is used essentially: it is what makes $\hat{\mu}_k$ a sufficient statistic for μ_Φ. Take a feature of the sample *set* that is not a function of an empirical measure, one reading the order in which samples arrived, say. It is not covered by $\mathcal{T}_k(\Phi) \le V^*(\Phi)$, though it is still covered by the universal bound $V^*(\nu)$, which is what licenses the paper's general claim that no label-free selector escapes identification by sampling more. We use no set-level feature. The answer-preserving hypothesis is used only to make $\mathsf{A}(\nu)$ readable from μ_Φ; it costs nothing here, because a feature that discards answer identity cannot drive a rule that must name an answer. Attainment is the only part that needs regularity, and for answer counts it is immediate: θ on a finite alphabet is identified by its empirical measure, and $a \mapsto \mathbb{P}[a^*{=}a \mid \theta]$ is continuous on the simplex under the a.s. no-tie hypothesis of Corollary 2.2.

Finally, the sampled-support convention is what aligns Theorem 2.1 with Theorem B.1, whose $T(\Phi)$ has always been a maximum over $\mathsf{A}(\Phi)$ (Appendix H), and with Theorem B.1(ii), whose shortlists have always been built from sampled candidates. Under the unrestricted convention these three would disagree, V^* could exceed Cov_k, and the limit in Theorem 2.1 would be false. Under it they are one object, and Appendix H prices the difference between the two conventions at no more than $1 - \mathrm{Cov}_k$.

C.2 Corollary 2.2

Take Φ the identity on answers, so $\mu_\Phi = \theta$. Fix a problem with $\theta_{(L)} > \theta_{(L+1)}$. The empirical frequencies converge a.s. to θ, so a.s. eventually the top-L answers by count are exactly the top-L by mass; hence $\mathbf{1}\{a^* \in \text{top-}L \text{ counts}\} \to \mathbf{1}\{R^* \le L\}$ a.s., and bounded convergence gives $S_L(k) \to \mathbb{P}[R^* \le L]$. For the chain of inequalities, both directions use the sampled-support convention. On the left, answering the plurality is one admissible count rule and in the limit it answers $a_{(1)} \in \mathsf{A}(\theta)$, so $\mathbb{P}[R^*{=}1] = \mathbb{E}[\mathbb{P}[a^*{=}a_{(1)} \mid \theta]] \le V^*(\text{counts})$. On the right, a maximum over $\mathsf{A}(\theta)$ is at most the sum over $\mathsf{A}(\theta)$,

$$\max_{a\in\mathsf{A}(\theta)} \mathbb{P}[a^*{=}a \mid \theta] \;\le\; \sum_{a\in\mathsf{A}(\theta)} \mathbb{P}[a^*{=}a \mid \theta] \;=\; \mathbb{P}[a^* \in \mathsf{A}(\theta) \mid \theta] \;=\; \mathbb{P}[R^*{<}\infty \mid \theta],$$

and taking expectations gives $V^*(\text{counts}) \le \mathbb{P}[R^* < \infty]$. This second step is exactly where the convention is needed: an unrestricted maximum may exceed the posterior mass on $\mathsf{A}(\theta)$, putting the ceiling above coverage, which is the same defect that would break attainment in Theorem 2.1. □

C.3 Proposition 2.3

Fix the budget k. A count rule acts by naming a position in the observed count ordering, so by Appendix H the best one answers $\arg\max_j \pi_j(\Phi)$ and attains $\mathcal{T}_k(\Phi) = \mathbb{E}_\Phi[\max_j \pi_j(\Phi)]$, while the plurality answers $j{=}1$ and attains $\mathbb{E}_\Phi[\pi_1(\Phi)]$, which is by definition its accuracy at budget k. Subtracting gives the displayed identity, which vanishes iff $\pi_1(\Phi) = \max_j \pi_j(\Phi)$ a.s. That condition

constrains position 1 alone, so rank calibration (monotonicity of π_j in j) implies it but is not implied by it; what monotonicity is equivalent to is the corresponding statement holding at *every* depth L, since "positions $1, \ldots, L$ are the L largest π_j for every L" says exactly that π_j is non-increasing (Corollary 2.4). Everything here is finite-k and refers to the *observed* ordering; by Corollary 2.2 the plurality's accuracy tends to $\mathbb{P}[R^*{=}1]$ and $\mathcal{T}_k$ to $V^*(\Phi)$, which is the limiting form quoted in the body.

For the sufficient condition, mass-calibration gives $\mathbb{P}[a^*{=}a_{(j)} \mid \theta] = \theta_{(j)}$, non-increasing in j because the $\theta_{(j)}$ are sorted; so given θ the Bayes answer is the highest-mass one, and since the observed count ordering converges a.s. to the mass ordering (Corollary 2.2), the plurality attains $V^*(\text{counts})$ in the limit. We do not claim exact finite-k optimality under mass calibration alone: at finite k the posterior over observed positions also depends on how the multinomial noise reorders near-tied masses. □

C.4 Corollary 2.4

Conditional on Φ, a shortlist naming a set J of observed count positions with $|J| \le L$ contains the truth with probability $\sum_{j \in J} \pi_j(\Phi)$, since the events "the truth occupies position j" are disjoint and exhaust the sampled candidates. The Bayes Φ-measurable L-shortlist therefore takes the L positions of largest $\pi_j(\Phi)$, and its value is $G_L(\Phi)$ because Φ determines the observed ordering, so positions and sampled answers name each other. Under rank calibration $\pi_j(\Phi)$ is non-increasing in j a.s., so those positions are $j = 1, \ldots, L$, whose value is $\mathbb{E}_\Phi[\sum_{j \le L} \pi_j(\Phi)] = \mathbb{P}[a^* \in \text{top-}L \text{ by count}] = S_L(k)$. The fixed-$L$ version needs only that the first L positions are a.s. the L largest $\pi_j(\Phi)$, which is what the argument uses. Taking $L = 1$ recovers Proposition 2.3, and letting $k \to \infty$ under Corollary 2.2 gives $G_L \to \mathbb{P}[R^* \le L]$ in the calibrated case. □

C.5 Theorem 2.6

This is the finite-k statement of Appendix K.5, restated for general L and both signs of the margin. The exact exponent is the Sanov rate (3) of whichever event is rare: I_L for $\{a^* \notin \text{top-}L\}$ when $m_L > 0$, and its mirror $\tilde{I}_L$, the rate of $\{a^* \in \text{top-}L\}$, when $m_L < 0$. The displayed sub-Gaussian bound is Hoeffding applied to the difference of two multinomial cell counts, with a union bound over the r wrong answers in the first case and over the L dominating wrong answers in the second, which is the $\max(r, L)$ prefactor. The proof is in Appendix K.5 and is not repeated.

D Count banks: construction and gate

Every cell is reduced to one object – per problem, a cluster id per sample plus the id of the correct cluster – and every downstream quantity is a function of that object alone.

Order of operations. (1) Extract a final answer from each completion using only the completion text. (2) Cluster the extracted strings with the project canonicaliser, which never reads a gold string. (3) *Only then* identify which cluster is correct, by verifying a cluster representative against the reference answer with `math_verify`. Step (3) scores a partition that is already fixed, so it cannot change what a selector sees; steps (1)–(2) are what must be, and are, label-free.

Why the release's own answer field is not used. The convenient path through the *Large Language Monkeys* release sets a sample's extracted answer to the gold string whenever the sample grades correct. That makes answer identity a function of the correctness label, which leaks gold into the very partition a label-free selector reads. We build the partition from sampled strings instead, by the order of operations above. The leak is not hypothetical: measured with that legacy extractor, at least one cluster mixes correct and incorrect samples on $104/128$ MATH/Llama-3-8B and $119/128$ MATH/Gemma-7B problems, so no count-based quantity computed on top of it is usable. Table 2 reports the gate on the rebuilt banks.

Head-limited closure. The audited canonicaliser resolves residual surface variation with an $O(m^2)$ `math_verify` closure over the m distinct keys of a problem, and declines above $m{=}64$. At $k{=}10^4$

Table 2: **Count-bank gate, per cell.** "unparsed" is the share of completions with no extractable final answer (none of them is labelled correct by the release grader); "regrade agree" is agreement between our `math_verify` re-grade and the release's `is_corrects`; "multi-truth" counts problems where more than one cluster verifies as correct and "truth mass split" the share of samples thereby left outside the named truth cluster. The last two columns recompute the headline limits with the truth cluster defined by the release's labels on the identical partition.

task	model	n	k	unparsed (%)	regrade agree (%)	multi-truth	truth mass split (%)	Δ vs. release labels $\Pr[R^*{=}1]$	$\Pr[R^*{\le}2]$
GSM8K	Llama-3-70B	127	10000	0.00	99.96	6	0.0303	+0.0000	+0.0000
GSM8K	Llama-3-8B	127	10000	0.00	99.53	37	0.3510	+0.0000	+0.0000
GSM8K	Qwen2.5-3B	1319	64	0.03	–	–	–	n/a	n/a
GSM8K	Qwen2.5-7B	1319	64	0.01	–	–	–	n/a	n/a
MATH-500	Qwen2.5-3B	500	64	0.03	–	–	–	n/a	n/a
MATH-500	Qwen2.5-7B	500	64	0.01	–	–	–	n/a	n/a
MATH	Gemma-2B	128	10000	12.68	99.40	109	0.2113	+0.0000	+0.0000
MATH	Gemma-7B	128	10000	9.07	98.94	115	0.2445	+0.0000	+0.0000
MATH	Llama-3-70B	128	10000	6.12	99.02	83	0.1295	+0.0000	+0.0000
MATH	Llama-3-8B	128	10000	6.25	99.14	94	0.1327	+0.0000	+0.0000

a MATH problem shows several hundred distinct keys, so that path would skip the closure almost everywhere. We instead (i) key any answer that parses to a number by its rounded value, which merges 1.5 with $\frac{3}{2}$ for free while keeping 2 and $2k{+}2$ apart, and (ii) run the exact closure over keys holding at least 3 samples, capped at 32. Both directions are conservative: an unmerged pair of spellings *splits* mass, which lowers measured concentration and lowers the truth's apparent rank-1 share, working against this paper's claims rather than for them.

Re-grading, and the sensitivity check. We name the truth cluster by re-grading rather than by inheriting the release's `is_corrects`, because that grader rejects correct answers carrying a unit ("108 m", "1.5 seconds") and the rejections would surface as clusters mixing correct and incorrect samples. The two graders agree on 98.9–99.96% of samples per cell. Because re-grading could shift the level of every curve, we recompute the rank limits with the truth cluster defined the release's way on the identical partition: the headline limits $\mathbb{P}[R^*{=}1]$ and $\mathbb{P}[R^*{\le}2]$ then move by $+0.0000$ on *every* pool cell, i.e. they are unchanged to four decimal places (both statistics are the last three columns of Table 2). Whatever the re-grading fixes, it is not what produces the result.

Two oracles for the truth cluster, and why both. Our re-grade caps how many clusters may receive a `math_verify` *equivalence* call (the largest 8; a symbolic key, which is cheaper, runs on the largest 256), because that call can hit SymPy's comparison timeout and a MATH problem at $k{=}10^4$ can carry thousands of clusters. A cap can only *miss* a correct cluster, which under-measures coverage, and under-measured coverage shrinks the denominator of the one-bit share, the single direction in which a shortcut here would flatter the paper rather than penalise it. We therefore take the union of two independent oracles: our re-grade, and the release's own per-sample labels (a cluster counts if a majority of its samples are release-correct). On MATH/Llama-3-8B the union recovers the truth on 5 problems the capped re-grade alone missed, reducing problems with no correct sample from 7 to 2. Both oracles only ever *score* the fixed, label-free partition.

Split truth clusters. On the MATH cells many problems carry more than one cluster that verifies as correct, a rare alternative spelling of the same answer. We name the largest as the truth and do *not* merge them, because merging clusters on the grounds that both match gold would use the gold answer to build the partition, which is exactly the leak ruled out above. A label-free selector faces the same split, so this is the honest accounting; the mass it costs is reported per cell and is at most a few parts in 10^3.

The prediction test's sampling model. In §4 the measured side subsamples the held-out half *without* replacement while the predicted side draws i.i.d. from $\hat{\theta}$, so the two differ by a finite-population correction. We cap the budget at an eighth of the held-out half, where that factor is at most $1/8$ of the (already small) sampling variance; it cannot account for agreement at the observed median absolute error.

E THE BIT PRICE WITHIN DIFFICULTY STRATA

Every bit share in this appendix is the count-ranked quantity of Corollary 2.5*(ii)*, computed from $\mathbb{P}[R^* \leq 2^B]$: it prices the top-2^B-by-count realization, never exceeds the Bayes capacity $G_{2^B}(\Phi_{\text{counts}})$, and equals it under Corollary 2.4. Read at cell level, the share of the coverage–selection gap closed by one bit separates by benchmark: 0.60–0.75 on the GSM8K cells against 0.13–0.49 on the MATH cells, with no overlap. The obvious alternative explanation is difficulty: GSM8K is easier than MATH, so "benchmark" could be standing in for pass@1, and the cross-cell comparison has only ten points to argue with.

Table 3 answers this with the per-problem data instead. Each cell is split into five equal-count bins of per-problem pass@1, and the one-bit share is recomputed within each bin. A bin is reported only where its coverage–selection gap is non-degenerate; "–" marks a bin on which self-consistency already equals coverage, so there is no gap to close and no share to define. Gold is used only to form the bins and to measure; no selector sees the stratification.

Two readings survive the table, and they are the ones §5 makes. *Within a cell, difficulty orders the price.* On each of the six cells with more than one graded bin the share is non-decreasing in the bin's pass@1, from 0.00 on the hardest MATH bins to 1.00 on bins with pass@1 between 0.28 and 0.49. *Across cells at matched difficulty, the benchmark ordering does not persist.* The graded strata of the two task families overlap in estimated pass@1 over roughly 0.26–0.50; there the MATH strata close 0.75–1.00 of their gap against 0.60–0.74 for the GSM8K strata, above them rather than below. The cell-level separation is therefore a statement about difficulty composition: a MATH cell averages over a hard tail on which the price is prohibitive, and a GSM8K cell has almost no problems in that regime at all (every GSM8K cell contributes exactly one graded bin, its hardest fifth).

Two bounds on the reading. The matched comparison rests on the overlap region, where shares still span 0.60–1.00, so difficulty organizes the price without being its only determinant; and pass@1 is estimated from the same samples that define R^*, so this table's per-bin levels are biased. §E.1 removes that second bound by re-estimating difficulty out of sample rather than arguing that the bias cancels. What transfers between cells is the stratum-level price, not a cell-level number.

Unextractable samples. Between 0.0% and 12.7% of pool completions carry no extractable final answer, the rate rising with benchmark difficulty and falling to 0.03% on the grid cells (Table 2); inspection shows these are truncated or degenerate generations (repetition loops), and *none* of them is labelled correct by the release grader. They are dropped from counts and their rate is reported per cell, rather than being silently imputed.

E.1 CROSS-FITTING THE DIFFICULTY ESTIMATE

The stratification above uses one sample bank twice: a problem's pass@1 places it in a bin, and the same samples give the rank R^* that prices the bin. The two uses share their noise. A problem whose truth was over-sampled by chance lands in an easier bin *and* gets a better rank there, so the price could look ordered by difficulty even if it were not. Saying the bias falls on both sides of a matched comparison is an argument; this section replaces it with a measurement.

The estimator. Split every problem's samples into two disjoint halves at random. Estimate pass@1 on the first half and cut the equal-count bins there; measure coverage, R^* and the bit price on the second. Reverse the halves and pool the two folds, so each problem contributes once per fold and no sample ever both places a problem in a stratum and prices it. Ties in half-sample pass@1, common at $k{=}64$ where the estimate moves in steps of $2/k$, are broken at random rather than by bank order. The pass@1 reported per bin is the *held-out* half's, which is unbiased for the stratum.

The control that isolates the circularity. Cross-fitting at $k/2$ also halves the sample budget, which moves levels on its own. The comparison in Table 4 is therefore against a same-sample control at the *same* sample size: bin and price on one half, folds reversed and pooled the same way. The two estimators differ only in whether the binning is independent of the measurement. Comparing the cross-fitted numbers against the published ones at full k would confound de-circularization with a smaller budget.

Table 3: Count-ranked one-bit share (Corollary 2.5*(ii)*) within equal-count per-problem difficulty bins. "bin 1" is the hardest fifth. A "–" marks a bin whose coverage–selection gap is degenerate (self-consistency already equals coverage), leaving no share to define. Within every cell the share is non-decreasing in the bin's pass@1; across cells, the strata that overlap in pass@1 do not preserve the benchmark ordering.

task	model		bin 1	bin 2	bin 3	bin 4	bin 5
GSM8K	Llama-3-8B	pass@1	0.26	0.71	0.91	0.98	1.00
		1-bit share	0.73	–	–	–	–
GSM8K	Llama-3-70B	pass@1	0.68	0.98	1.00	1.00	1.00
		1-bit share	0.75	–	–	–	–
MATH	Llama-3-8B	pass@1	0.00	0.04	0.15	0.41	0.82
		1-bit share	0.00	0.04	0.36	1.00	–
MATH	Llama-3-70B	pass@1	0.02	0.15	0.49	0.85	0.99
		1-bit share	0.12	0.56	1.00	–	–
MATH	Gemma-7B	pass@1	0.00	0.03	0.09	0.28	0.69
		1-bit share	0.00	0.15	0.29	1.00	–
MATH	Gemma-2B	pass@1	0.00	0.01	0.03	0.09	0.39
		1-bit share	0.00	0.00	0.04	0.44	1.00
GSM8K	Qwen2.5-3B	pass@1	0.33	0.84	0.96	1.00	1.00
		1-bit share	0.60	–	–	–	–
GSM8K	Qwen2.5-7B	pass@1	0.50	0.96	1.00	1.00	1.00
		1-bit share	0.74	–	–	–	–
MATH-500	Qwen2.5-3B	pass@1	0.03	0.31	0.73	0.94	1.00
		1-bit share	0.11	0.75	–	–	–
MATH-500	Qwen2.5-7B	pass@1	0.06	0.46	0.85	0.98	1.00
		1-bit share	0.31	1.00	–	–	–

What it changes: almost nothing. Across the strata both estimators grade, the median shift in the one-bit share is 0.00 and the largest is 0.13 (0.22 over 20 independent splits); the largest shift in a bin's pass@1 level is 0.014. The one stratum that moves materially is MATH-500/Qwen2.5-3B's hardest fifth, where the same-sample estimate reads 0.06 and the cross-fitted one 0.15: the circularity was *under*-stating the price there, not inflating it. Both readings of §5 survive. Within cells the price is non-decreasing in the bin's held-out pass@1 on 6 of 6 cells with more than one graded stratum, in 20 of 20 splits. Across cells at matched difficulty (pass@1 in 0.26–0.49) the MATH strata close 0.60–1.00 of their gap against 0.61–0.72 for the GSM8K strata. The cell-level ordering, which has MATH strictly below GSM8K with no overlap, does not return at matched difficulty: the two bands overlap in 20 of 20 splits, and the MATH band reaches above the GSM8K band in 20 of 20. Over the 11 bridgeable strata one bit closes 0.37–1.00 of the gap, two bits 0.79–1.00 and five bits 1.00, against 0.36–1.00, 0.77–1.00 and 1.00 same-sample; the median rank of the truth on failures is 2.0–3.0, and it is exactly 2 on 82% of them. The hard tail stays a tail: on the strata with pass@1 below 0.05 one bit closes 0.00–0.15 and the median rank on failures is 5–89.

Two further notes. A stratum is graded only when its coverage–selection gap spans at least five problems; a gap of one or two problems makes the share a ratio of small integers that one problem moves from 0 to 1, and quoting it beside strata built on hundreds would read as a measurement rather than as rounding. This costs one stratum relative to Table 3 and is applied to both estimators. Second, the same cross-fit answers the stratified reading of the grid decomposition (Fig. 13): with difficulty estimated out of sample, the gap divided by coverage is non-increasing in the bin's held-out pass@1 on 10 of 10 cells, in 20 of 20 splits. The bins there are the count banks', not that figure's equal-width bins of $1 - \text{pass@1}$, so this confirms the monotone trend the figure reads rather than reproducing its levels.

Table 4: **The bit price with difficulty estimated out of sample.** Per-problem samples are split in two: bins are cut on one half, the price is measured on the other, and the folds are reversed and pooled. "pass@1 (held out)" is the measured half's, so it is unbiased for the stratum. The same-sample row is the circular estimator *at the same sample size*, which is what isolates the circularity from the halved budget; brackets are 95% bootstrap intervals over problems. "–" marks a stratum whose gap is degenerate or spans fewer than five problems. Every reading of Table 3 survives: the price is monotone in difficulty within cells, and the benchmark ordering does not return at matched difficulty.

task	model		bin 1	bin 2	bin 3	bin 4	bin 5
GSM8K	Llama-3-8B	pass@1 (held out)	0.26	0.71	0.91	0.98	1.00
		1-bit, same-sample	0.72	–	–	–	–
		1-bit, cross-fitted	0.72 [0.48,0.94]	–	–	–	–
GSM8K	Llama-3-70B	pass@1 (held out)	0.68	0.98	1.00	1.00	1.00
		1-bit, same-sample	0.75	–	–	–	–
		1-bit, cross-fitted	0.75 [0.00,1.00]	–	–	–	–
MATH	Llama-3-8B	pass@1 (held out)	0.00	0.04	0.15	0.41	0.82
		1-bit, same-sample	0.00	0.04	0.37	1.00	–
		1-bit, cross-fitted	0.00 [0.00,0.00]	0.04 [0.00,0.13]	0.37 [0.14,0.61]	1.00 [1.00,1.00]	–
MATH	Llama-3-70B	pass@1 (held out)	0.02	0.15	0.49	0.85	0.99
		1-bit, same-sample	0.10	0.50	1.00	–	–
		1-bit, cross-fitted	0.10 [0.00,0.24]	0.50 [0.28,0.72]	1.00 [1.00,1.00]	–	–
MATH	Gemma-7B	pass@1 (held out)	0.00	0.03	0.09	0.28	0.69
		1-bit, same-sample	0.00	0.12	0.26	1.00	–
		1-bit, cross-fitted	0.00 [0.00,0.00]	0.12 [0.02,0.26]	0.26 [0.09,0.46]	1.00 [1.00,1.00]	–
MATH	Gemma-2B	pass@1 (held out)	0.00	0.01	0.03	0.09	0.39
		1-bit, same-sample	0.00	0.00	0.04	0.43	–
		1-bit, cross-fitted	0.00 [0.00,0.00]	0.00 [0.00,0.00]	0.04 [0.00,0.13]	0.43 [0.21,0.64]	–
GSM8K	Qwen2.5-3B	pass@1 (held out)	0.35	0.85	0.96	0.99	0.99
		1-bit, same-sample	0.61	–	–	–	–
		1-bit, cross-fitted	0.61 [0.53,0.69]	–	–	–	–
GSM8K	Qwen2.5-7B	pass@1 (held out)	0.51	0.96	1.00	1.00	1.00
		1-bit, same-sample	0.76	–	–	–	–
		1-bit, cross-fitted	0.76 [0.66,0.85]	–	–	–	–
MATH-500	Qwen2.5-3B	pass@1 (held out)	0.04	0.31	0.73	0.93	0.99
		1-bit, same-sample	0.06	0.73	–	–	–
		1-bit, cross-fitted	0.15 [0.07,0.24]	0.60 [0.49,0.71]	–	–	–
MATH-500	Qwen2.5-7B	pass@1 (held out)	0.07	0.47	0.85	0.98	0.99
		1-bit, same-sample	0.32	0.94	–	–	–
		1-bit, cross-fitted	0.32 [0.22,0.42]	0.90 [0.80,0.98]	–	–	–

F STATEMENTS DEFERRED FROM THE MAIN TEXT

The body states the paper's spine: the guessing curve (Theorem B.1), the rank zero–one law (Theorem B.2), and the margin law (Theorem B.3). This appendix collects the statements the body cites but does not display, so that the shared proofs refer to numbered statements.

F.1 CONVENTIONS AND THE STATIC CURVE

The chain (1) splits the coverage–selection gap into three measurable levels, written with $T_{\mathrm{tr}} := T(\Phi_{\mathrm{traces}})$ and $T_{\mathrm{ct}} := T(\Phi_{\mathrm{counts}})$:

$$\underbrace{\mathrm{Cov}_k - F}_{\text{the gap}} = \underbrace{(\mathrm{Cov}_k - T_{\mathrm{tr}})}_{\text{label-free irreducible}} + \underbrace{(T_{\mathrm{tr}} - T_{\mathrm{ct}})}_{\text{trace-recoverable}} + \underbrace{(T_{\mathrm{ct}} - F)}_{\text{count-recoverable}} . \tag{2}$$

Cov_k, F and (via rank rules) T_{ct} are estimable directly; T_{tr} is not computable in closed form, so the real-data decomposition is deliberately partial: count level measured, trace level bracketed below by the best selector tested and above by $\mathrm{Cov}_k - T_{\mathrm{ct}}$.

F.2 RANK SELECTION

Self-consistency picks the modal answer, and the multiset of answers is all a count-only selector sees, so one asks whether plurality is the Bayes rule for it. The answer is conditional on the following dominance hypothesis, whose second clause says that the more frequent of two answers is, in likelihood, the more plausibly correct one. It holds, for example, whenever $\theta_{a^*} \geq \max_{a \neq a^*} \theta_a$ almost surely under the prior.

Assumption 2 (Answer-label symmetry with correct-count dominance)**.** The joint prior law of (a^*, θ) is invariant under permutations of the answer alphabet, and for every pair $a \neq b$ and every count vector n with $n_a > n_b$, $\mathbb{P}[N = n \mid a^* = a] \geq \mathbb{P}[N = n \mid a^* = b]$.

Proposition F.1 (Rank-1 optimality; when plurality is the count ceiling)**.** *Under Assumptions 1 and 2's answer-label symmetry, there exists a Bayes-optimal count-only selector that is a* rank rule*: a map from the sorted count profile to a count-rank. Plurality is the constant rank-1 rule. It maximizes $\mathbb{P}[g(N) = a^*]$ over all count-only g, with accuracy F, if and only if the answer selected by the fixed plurality tie rule maximizes $\mathbb{P}[a^* = a \mid N = n]$ on every positive-probability profile n. The dominance clause of Assumption 2 is sufficient for that condition but not necessary; where it fails on a positive-probability profile, a Bayes rank rule strictly improves on plurality and $T(\Phi_{\text{counts}}) > F$. Proof in Appendix J.*

Two consequences. *(a)* How much of the gap is count-recoverable is at most $\mathbb{P}[a^* \text{ at rank } 2, 3, \cdots \mid \text{covered}]$, the mass a rank rule competes for; mass at deep ranks means the information is genuinely lost to the counts, while mass at rank 2 means it is present but capturable only if the sorted profile *distinguishes* those problems from rank-1 ones, since a rule that always answers rank 2 forfeits every problem plurality gets right. *(b)* A calibrated rank rule, not plurality, is therefore the honest baseline for any trace-aware method; on our cells the two coincide numerically, which is itself the finding of §6.

Remark F.2 (Where the truth goes: the runner-up phenomenon). When a single shared misconception dominates the errors (the two-mode regime of §2), the truth is out-emitted by exactly one wrong answer, so it sits at count-*rank 2* almost surely, and a label-free "pick the runner-up" rule recovers it. Our synthetic sweep makes this quantitative (§2, Fig. 9): at $p = 0.30$, $k = 64$, $\rho = 0.9$ the truth is at rank 2 with probability 0.999, plurality scores 0.0015, and the runner-up rule scores 0.9985. Genuine count-only collapse requires a different structure, *two* answer modes of equal mass, where the sorted profile cannot say which is correct; that is the attained-ceiling family of §2 (Theorem F.4), not the single-dominant regime. The operative count-only ceiling is therefore $T(\Phi_{\text{counts}})$, the best rank rule, which we measure alongside F; the gap between them is the count-recoverable term of (2). Systematic string-level asymmetries (format, length bias) are a separate, trace-level resource and are out of the count-only class by construction.

F.3 Exact exponents

The exponent that governs the rank zero–one law is not the Hoeffding quantity $m_L^2/2$ but an exact large-deviation rate.

Because the count vector is multinomial, $N \sim \text{Mult}(k, \theta)$, the event $\{a^* \notin \text{top-}L\}$ is a finite union, over which set S of L wrong answers weakly beats the truth, of the *convex* sets $\mathcal{C}_S = \{\nu : \nu_a \geq \nu_{a^*}\ \forall a \in S\}$. Sanov's theorem therefore gives its rate as a minimum of convex programs,

$$I_L \;=\; \min_{|S|=L} \; \inf_{\nu \in \mathcal{C}_S} \text{KL}(\nu \,\|\, \theta), \tag{3}$$

each with an explicit solution. For $L = 1$ the program has a closed form: at the optimum the two binding coordinates tie at the geometric mean $\sqrt{\theta_{a^*}\theta_b}/Z$ with $Z = 1 - \theta_{a^*} - \theta_b + 2\sqrt{\theta_{a^*}\theta_b}$, and the value is $-\log Z$. Writing

$$I(\theta_{a^*}, \theta_b) \;=\; -\log\!\big(1 - (\sqrt{\theta_{a^*}} - \sqrt{\theta_b})^2\big) \;\geq\; \tfrac{1}{2}(\theta_{a^*} - \theta_b)^2, \tag{4}$$

we get $I_1 = \min_{b \neq a^*} I(\theta_{a^*}, \theta_b)$: the Bhattacharyya-type Chernoff exponent, which dominates the Hoeffding exponent for every pair and strictly so unless the masses coincide.

The Gaussian law of Theorem B.3 is not uniform as $\sigma \to 0$, and that is not a technicality about a measure-zero corner: at $k = 64$ the hardest difficulty quintile has $kp = O(1)$, so the truth appears a handful of times, not $\Theta(k)$ times. That regime deserves its own limit rather than an exclusion.

Corollary F.3 (Threshold, and finite-k rates: sampling entrenches the loser)**.** *Under Assumption 1, with $q = (1-p)w$ and r the number of wrong answers of positive mass:* ***(i)*** *if $w < p/(1-p)$ (i.e. $m > 0$) then $\mathbb{P}[\text{plurality} = a^*] \to 1$, with $1 - \mathbb{P}[\text{plurality} = a^*] \leq r\, e^{-km^2/2}$;* ***(ii)*** *if $w > p/(1-p)$ (i.e. $m < 0$) then the plurality converges to the dominant wrong answer, and in the two-mode reduction*

$$\mathbb{P}[\text{plurality} = a^*] = \frac{C(p,q)}{\sqrt{k}}\, e^{-k\, I(p,q)}\,(1 + o(1)), \qquad C(p,q) = \frac{1}{(1 - \sqrt{p/q})\,\sigma_* \sqrt{2\pi}},$$

with $I(p,q)$ from (4) *and $\sigma_*^2 = 2\sqrt{pq}/(1-(\sqrt{p}-\sqrt{q})^2)$. Voting accuracy decays* exponentially in *k: every additional sample makes selection worse. The crossover is at k of order $1/m^2$. Proof in Appendix K.*

The lattice correction to the margin law. Because the contest walk of Theorem B.3 moves on the span-1 integer lattice, the Gaussian limit carries a one-term Edgeworth correction: a continuity part $-\phi(z)/(2\sigma\sqrt{k})$, together with a skewness part carrying the standardized third central moment γ_3 of the step law, which vanishes in the mean. The continuity part is negative at every z and z-free up to the $\phi(z)$ envelope, so the theory predicts that the held-out voting curve sits *below* Φ throughout, by an amount of order $k^{-1/2}$. That is a falsifiable prediction of the *sign* of the sub-Φ dip, derived from the lattice alone and owing nothing to estimation noise; §P.5 reports the test.

F.4 The attained ceiling, the trace escape, and the one-bit closure

Theorem F.4 (Attained no-free-lunch ceiling)**.** *There is an explicit task family with finite-k coverage $\mathrm{Cov}_k = 1-2^{-k} \uparrow 1$ on which $T(\Phi) = 1/2$ for* every *label-free feature Φ of the sample set. Construction: each problem has two answer modes, the truth and one wrong answer, with* equal *emission probability ($m=0$), and the map swapping their string identities preserves the joint law of the traces. Then the posterior $\mathbb{P}[a^* = a \mid \Phi]$ is symmetric in the two modes, $H_\infty(a^* \mid \Phi) = 1$ bit, and no selector, count- or trace-based, beats a coin. Proof in Appendix L.*

This is the exact sense in which the ceiling is fundamental rather than an artifact of weak selectors, and it corrects an over-statement tempting from the single-dominant regime, where the truth is merely out-voted but still identifiable at rank 2. It is a knife-edge in two senses. In m: the ceiling does not jump at $m{=}0$ but passes through $1/2$ with slope $\sqrt{k}\,\phi(0)/\sigma$, so a problem within $O(\sigma/\sqrt{k})$ of balance is, for selection purposes, already at the coin ceiling. And in trace symmetry, which Theorem F.4 imposes by fiat and real traces need not obey, Theorem F.5 turns dropping that symmetry into a rate.

Theorem F.5 (Trace separation: the exponential escape from the coin ceiling)**.** *On the balanced stratum of Theorem F.4, let $k' = n_u + n_v$ be the number of samples in the two live clusters and let Φ_{tr} be (counts, the k' trace statistics), whose per-sample law is P_c in the truth's cluster and P_w in the wrong mode's, with Bhattacharyya affinity $\rho = \mathrm{BC}(P_c, P_w) = \int \sqrt{dP_c\, dP_w}$. Then*

$$1 - \tfrac{1}{2}\,\rho^{\,k'} \;\;\le\;\; T(\Phi_{\mathrm{tr}}) \;\;\le\;\; \tfrac{1}{2}\big(1+\sqrt{1-\rho^{\,2k'}}\big).$$

*Consequently: **(i) (Dichotomy.)** $T(\Phi_{\mathrm{tr}}) = \frac{1}{2}$ for every k'* if and only if *$\rho = 1$, i.e. iff the statistic is mode-symmetric; the no-free-lunch ceiling is knife-edge in a second sense beyond $m = 0$. **(ii) (Exponential escape.)** If $\rho < 1$ then $1 - T(\Phi_{\mathrm{tr}}) \le \frac{1}{2}e^{-k'\log(1/\rho)}$. **(iii) (Additivity.)** Without balance, the posterior log-odds decompose as $\Lambda(\Phi_{\mathrm{tr}}) = \Lambda(\Phi_{\mathrm{ct}}) + \sum_{i\in u}\lambda(t_i) - \sum_{i\in v}\lambda(t_i)$ with $\lambda = \log dP_c/dP_w$: count and trace evidence add on the log-odds scale. **(iv) (Sharp exponent.)** The exact Bayes error exponent is the Chernoff information $E(n_u, n_v) = \max_{s\in[0,1]}[n_u d_s(P_c, P_w) + n_v d_s(P_w, P_c)]$, which is at least $k'\log(1/\rho)$ and equals it when $n_u = n_v$. Equality may also occur for likelihood-ratio-symmetric pairs, so imbalance need not strictly improve the exponent. Proof in Appendix L.2; verified independently of the proof by brute-force enumeration of the outcome space.*

Lemma F.6 (Pooling inflates the affinity)**.** BC *is jointly concave in its pair of arguments. Hence for any mixing weights π over problems,*

$$\mathrm{BC}\Big(\textstyle\sum_j \pi_j P_c^{(j)},\ \sum_j \pi_j P_w^{(j)}\Big) \;\ge\; \sum_j \pi_j\, \mathrm{BC}\big(P_c^{(j)}, P_w^{(j)}\big),$$

so an affinity estimated from pooled-across-problems samples estimates the affinity of the mixture *laws, which is at least the mean per-problem affinity. Heterogeneity therefore biases $\hat{\rho}$* toward 1. *Proof in Appendix L.4.*

Theorem F.7 (Top-two reduction and the list law)**.** *Let $G_L(\Phi_{\mathrm{counts}})$ be the count-only Bayes list ceiling, and let $S_L = \mathbb{P}[a^* \in \mathrm{Top}_L^{\mathrm{count}}]$ be the fixed top-L-by-count success rate. If the second-largest wrong emission mass satisfies $p > (1-p)w_2$ (the truth out-emits all but at most one wrong answer), then $S_2 \to \mathrm{Cov}_k$, and since $S_2 \le G_2(\Phi_{\mathrm{counts}}) \le \mathrm{Cov}_k$ the Bayes list ceiling is squeezed to the same limit, $G_2(\Phi_{\mathrm{counts}}) \to \mathrm{Cov}_k$ as $k \to \infty$. On the symmetric stratum of Theorem F.4, both the fixed and Bayes curves move from $1/2$ at $L=1$ to $\mathrm{Cov}_k = 1-2^{-k}$ at $L=2$. Proof in Appendix K.*

Proposition F.8 (One bit per problem closes the symmetric stratum)**.** *On the symmetric stratum (mass Δ of problems with $T(\Phi) = 1/2$), one perfect binary comparison closes the finite-k selection-to-coverage gap, attaining $\mathrm{Cov}_k = 1 - 2^{-k}$; for $k \geq 2$, a zero-bit transcript cannot attain that value. As $k \to \infty$, the achieved accuracy tends to one without repairing the generation misses at any finite k.*

Proposition F.9 (Asymptotic gap identity)**.** *Under Assumption 1, suppose the prior almost surely gives a non-degenerate answer distribution in which at most one wrong answer attains the maximal wrong mass. Then, as $k \to \infty$,*

$$\mathrm{Cov}_k \to \mathbb{P}[p > 0], \qquad F_k \to \mathbb{P}[m > 0] + \tfrac{1}{2}\,\mathbb{P}[m = 0,\ p > 0],$$

and hence

$$\lim_{k\to\infty} (\mathrm{Cov}_k - F_k) \;=\; \mathbb{P}[\,p > 0,\ m < 0\,] \;+\; \tfrac{1}{2}\,\mathbb{P}[\,p > 0,\ m = 0\,] :$$

the asymptotic coverage–selection gap equals *the prior mass of the covered, margin-negative stratum (plus half the boundary). Proof in Appendix K.4.*

G Position relative to classical decision theory

This appendix relates our formulation to classical decision theory and identifies the corresponding prior results.

Per realization of Φ, $L \mapsto \sum_{i \leq L} p_{(i)}$ is the majorization profile (Lorenz curve) of the posterior (Marshall et al., 2011). Comparing experiments by their curves is Blackwell's comparison restricted to the 0–1 list-loss class, an instance of his own weaker k-decision comparisons (Blackwell, 1953); Sherman's theorem bridges that ordering to Hardy–Littlewood–Pólya majorization (Sherman, 1951), with Torgersen (1991) for the full edifice, and the template "decision class $\to$ single information functional" is likewise established (Cabrales et al., 2013). Within this canon, G_L is the success probability of the optimal list-of-L decoder (Forney, 1968), the bit reading is the depth-B query tree, and $G_1 = 2^{-H_\infty}$ anchors the guessing lineage (Massey, 1994; Arıkan, 1996). The bit-budgeted reading belongs to the *rate-limited guessing* line, guessing with a helper constrained to a finite message (Weinberger & Shayevitz, 2020; Arıkan & Merhav, 1998), which studies guessing *moments* and asymptotic rates, where Theorem B.1(iii) is a single-shot, success-probability specialization we prove directly.

The contribution specific to the present setting lies in the sampling dynamics: the finite-k margin law (Theorem B.3) and the rank zero–one law (Theorem B.2), together with the support-size-only converse of the verification value, which binds even omniscient helpers, and the curve's empirical program (§3). The classical large-sample comparison literature (Moscarini & Smith, 2002; Mu et al., 2021) ranks *two fixed experiments* under n i.i.d. repetitions with Chernoff- and Rényi-divergence rates, and our sharp exponent in §2 is a quantity of that family; to our knowledge it does not describe the *value trajectory* of the single growing self-consistency experiment or the per-problem step limits at the truth's mass-rank, which is where this paper's theoretical contribution is concentrated.

H Proof of Theorem B.1 (ceiling identity)

A label-free selector is a Markov kernel $g(\cdot \mid \Phi)$ into the sampled answers $\mathsf{A}(\Phi)$. Its accuracy is $\mathbb{P}[g = a^*] = \mathbb{E}_\Phi[\sum_{a \in \mathsf{A}(\Phi)} g(a \mid \Phi)\,\mathbb{P}[a^* = a \mid \Phi]]$, maximized pointwise by the sampled candidate of largest posterior mass. Thus $T(\Phi) = \mathbb{E}_\Phi[\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi]] \leq 2^{-H_\infty(a^* \mid \Phi)}$. The restriction to $\mathsf{A}(\Phi)$ is the sampled-support convention of Theorem 2.1 and is forced by the range of g. The inequality is an equality whenever an unrestricted posterior maximizer is sampled almost surely. If the unrestricted Bayes rule gains over the restricted rule, it must select an unsampled answer and can be correct only when the truth is unsampled; hence its advantage is at most $1 - \mathrm{Cov}_k$, which bounds the whole discrepancy between the two conventions. For the chain (1): if $\Psi = f(\Phi)$ is a deterministic coarsening that preserves the sampled answer set, so $\mathsf{A}(\Psi) = \mathsf{A}(\Phi)$, then that common set is Ψ-measurable and the tower rule followed by Jensen on the fixed index set gives $\max_{a \in \mathsf{A}(\Psi)} \mathbb{P}[a^* = a \mid \Psi] = \max_{a \in \mathsf{A}(\Psi)} \mathbb{E}[\mathbb{P}[a^* = a \mid \Phi] \mid \Psi] \leq \mathbb{E}[\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi] \mid \Psi]$; taking expectations gives $T(\Psi) \leq T(\Phi)$. Apply with traces $\to$ counts, which does preserve the

sampled answer set. The last step of (1), $T(\Phi_{\text{counts}}) \geq F$, is not an instance of this and does not need to be: answering the plurality is one admissible count rule, so its accuracy is at most the best one's. Coverage dominates because a selector returns a sampled answer, so $\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi] \leq \sum_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi] = \mathbb{P}[a^* \in \mathsf{A}(\Phi) \mid \Phi]$, whose expectation is Cov_k. □

I Proof of Theorem B.1 (the guessing curve)

Throughout, posteriors are over the sampled candidates given Φ, and $p_{(1)} \geq p_{(2)} \geq \cdots$ are their sorted masses.

(i) Endpoints. $G_1 = \mathbb{E}_\Phi[p_{(1)}] = \mathbb{E}_\Phi[\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi]] = T(\Phi)$ by Theorem B.1 (with the sampled-support convention of Theorem B.1(i)). For L at least the number of sampled candidates, $\sum_{i \leq L} p_{(i)} = \mathbb{P}[a^* \in \text{sampled set} \mid \Phi]$, whose expectation is exactly the coverage probability; monotonicity in L is clear.

(ii) Lists. An L-list rule from Φ is a Φ-measurable set $S(\Phi)$ with $|S| \leq L$; its success probability is $\mathbb{E}_\Phi[\mathbb{P}[a^* \in S(\Phi) \mid \Phi]]$, maximized pointwise by the L highest-posterior candidates, giving G_L.

(iii) Verification, exact. We prove the statement for a transcript with at most M values and only then specialize to $M = 2^B$, which avoids any question about shortlist sizes that are not powers of two. *Achievability:* the transcript names which of the top-M posterior candidates is correct, so accuracy $= \mathbb{E}_\Phi[\sum_{i \leq M} p_{(i)}] = G_M$; when $M = 2^B$ that name is B bits, deliverable as B membership queries bisecting the shortlist. *Converse:* model any protocol as a transcript V jointly distributed with (a^*, Φ), $|\mathrm{range}(V)| \leq M$ (adaptivity and shared randomness absorbed into the joint law). The Bayes selector given (Φ, V) achieves

$$\mathbb{E}_\Phi\Big[\sum_v \mathbb{P}[V = v \mid \Phi] \max_a \mathbb{P}[a^* = a \mid \Phi, V = v]\Big] = \mathbb{E}_\Phi\Big[\sum_v \max_a \mathbb{P}[a^* = a,\ V = v \mid \Phi]\Big].$$

Fix Φ and let $a(v)$ be a maximizer for transcript value v. Grouping the v's by their maximizer,

$$\sum_v \mathbb{P}[a^* = a(v),\, V = v \mid \Phi] = \sum_{a \in \mathcal{A}^*} \sum_{v:\, a(v) = a} \mathbb{P}[a^* = a,\, V = v \mid \Phi] \ \leq \ \sum_{a \in \mathcal{A}^*} \mathbb{P}[a^* = a \mid \Phi],$$

where $\mathcal{A}^*$ is the set of distinct maximizers, $|\mathcal{A}^*| \leq M$. The right side is at most $\sum_{i \leq M} p_{(i)}$. Taking $\mathbb{E}_\Phi$ gives accuracy $\leq G_M$, matching achievability; with $M = 2^B$ this is Corollary 2.5*(i)*. (Note this is *stronger* than the chain-rule bound MG_1, which follows from it by $p_{(i)} \leq p_{(1)}$.) The maximizing shortlist is the M highest-posterior sampled candidates, which need not be the M highest by count; the count-ranked channel of Corollary 2.5*(ii)* therefore attains only $S_M(k) \leq G_M$ in general, with equality under Corollary 2.4.

(iv) Data processing. Let $a^* \to \Phi \to \Psi$ be a Markov chain (deterministic $\Psi = f(\Phi)$ is the special case). For any Ψ-measurable S with $|S| \leq L$, conditional independence gives $\mathbb{P}[a^* \in S \mid \Psi] = \mathbb{E}[\mathbb{P}[a^* \in S \mid \Phi] \mid \Psi] \leq \mathbb{E}[\max_{|S'| \leq L} \mathbb{P}[a^* \in S' \mid \Phi] \mid \Psi]$; maximize over S and take expectations. (The Markov form is what the k-monotonicity argument uses: given the k-sample counts, deleting a uniformly random sample decrements answer a with probability N_a/k regardless of a^*, so $(k-1)$-counts are a stochastic garbling of k-counts, and by exchangeability their law equals that of a fresh $(k-1)$-sample count vector.)

(v) Concavity. $G_{L+1} - G_L = \mathbb{E}[p_{(L+1)}]$, nonincreasing in L because the masses are sorted; with $G_0 = 0$, concavity gives $G_L \leq L\, G_1$ and hence $G_{2^B} \leq \min(1, 2^B G_1)$. □

J Proof of Proposition F.1 (rank-1 optimality)

By Appendix H a Bayes count-only selector outputs $\arg\max_a \mathbb{P}[a^* = a \mid N]$. Under label symmetry (Assumption 2) the prior is uniform over answer identities, so $\mathbb{P}[a^* = a \mid N = n] \propto \mathbb{P}[N = n \mid a^* = a]$, which by exchangeability of the sampling depends on a only through its count n_a and the sorted profile of n. Hence the Bayes selector is a function of (count-of-a, sorted profile), so a Bayes-optimal *rank rule* exists. With the tie rule fixed, plurality is Bayes-optimal exactly when its selected maximal-count answer belongs to $\arg\max_a \mathbb{P}[a^* = a \mid N = n]$ for every n in the support.

If this fails on a positive-probability profile, replacing plurality there by a posterior maximizer gives a strict improvement.

Pairwise dominance is a convenient sufficient condition. Fix a, b with $n_a > n_b$; then

$$\frac{\mathbb{P}[a^* = a \mid N = n]}{\mathbb{P}[a^* = b \mid N = n]} = \frac{\mathbb{P}[N = n \mid a^* = a]}{\mathbb{P}[N = n \mid a^* = b]},$$

which is ≥ 1 for all such pairs under dominance. Therefore no lower-count answer has larger posterior than a maximal-count answer, and rank 1 is Bayes-optimal. The converse does not follow: dominance can fail between two non-maximal answers while a maximal-count answer remains the posterior maximizer. □

Sufficient condition for the dominance clause. Suppose the prior draws an unordered probability vector and assigns a^* to a coordinate whose value is almost surely maximal ($\theta_{a^*} \geq \max_{a \neq a^*} \theta_a$ a.s.). Conditional on the unordered vector, exchanging the roles of a and b maps the event $\{N = n, a^* = a\}$ to $\{N = n', a^* = b\}$ with n' the swap of n_a, n_b; i.i.d. sampling (Assumption 1) then gives

$$\frac{\mathbb{P}[N = n \mid a^* = a]}{\mathbb{P}[N = n \mid a^* = b]} = \mathbb{E}\left[\left(\tfrac{\theta_a}{\theta_b}\right)^{n_a - n_b} \,\middle|\, \cdot\right] \;\geq\; 1 \quad \text{when } n_a > n_b \text{ and } \theta_a \geq \theta_b \text{ a.s. under } a^* = a,$$

where the expectation is over the conditional law of θ given $a^* = a$ and the unordered vector. The correlated-error regime $\theta_W > \theta_{a^*}$ violates exactly this hypothesis; see Remark F.2.

K Proofs for Section 2

Throughout, conditional on the problem, the k answers are i.i.d. with $\mathbb{P}[a^*] = p$, $\mathbb{P}[W] = (1-p)w$, and $\mathbb{P}[a] = (1-p)q_a$ for the other wrong answers, over a finite alphabet.

K.1 Corollary F.3: the asymptotic threshold

By the strong law of large numbers, $N_a/k \to \theta_a$ a.s. for every a in the (finite) alphabet, where $\theta_{a^*} = p$ and $\theta_a = (1-p)q_a$ otherwise.

(i) If $w < p/(1-p)$ then $\theta_{a^*} > \max_{a \neq a^*} \theta_a$; a.s. for all large k the count of a^* strictly exceeds every other count, so the plurality is a^* eventually, and by bounded convergence $\mathbb{P}[\text{plurality} = a^*] \to 1$.

(ii) If $w > p/(1-p)$ then $\theta_W = (1-p)w > p = \theta_{a^*}$, so a.s. for all large k, $N_W > N_{a^*}$ and the plurality is not a^*; hence $\mathbb{P}[\text{plurality} = a^*] \to 0$. (At equality the margin is driven by CLT fluctuations and the limit is nondegenerate; we do not use this case.) □

K.2 Corollary F.3: the finite-k bounds

Let $X_i = \mathbf{1}\{\text{sample } i = a^*\} - \mathbf{1}\{\text{sample } i = W\} \in \{-1, 0, 1\}$, i.i.d. with mean $m = p - (1-p)w$, and $S_k = \sum_{i=1}^k X_i = N_{a^*} - N_W$.

(i) If $m > 0$: $\{\text{plurality} \neq a^*\}$ requires $N_a \geq N_{a^*}$ for some wrong answer a; each such answer has per-sample margin $p - (1-p)q_a \geq p - (1-p)w = m$, so by Hoeffding for i.i.d. variables with range $[-1, 1]$ and a union bound over the r wrong answers with positive mass, $\mathbb{P}[\text{plurality} \neq a^*] \leq r\, e^{-km^2/2}$. (Ties are absorbed into the failure event, so the bound is conservative.)

(ii) If $m < 0$: $\{\text{plurality} = a^*\}$ requires $N_{a^*} \geq N_W$, i.e. $S_k \geq 0$. By the same Hoeffding bound, $\mathbb{P}[S_k \geq 0] = \mathbb{P}[S_k - km \geq -km] \leq \exp(-2(km)^2/(4k)) = e^{-km^2/2}$; the sharp prefactor for this case is derived in the Bahadur–Rao computation below. □

K.3 Theorem B.3

With X_i as above, plurality $= a^*$ on the two-mode contest iff $S_k = N_{a^*} - N_W > 0$. The X_i are i.i.d., bounded, with mean m and variance $\sigma^2 = \mathbb{E}X_i^2 - m^2 = (p + (1-p)w) - m^2$. By the Berry–Esseen theorem, $\mathbb{P}[S_k > 0] = \mathbb{P}\big(\frac{S_k - km}{\sigma\sqrt{k}} > -\frac{\sqrt{k}\,m}{\sigma}\big) = \Phi(\sqrt{k}\,m/\sigma) + R_k$, with $|R_k| \leq$

$C\,\mathbb{E}|X_1 - m|^3/(\sigma^3\sqrt{k})$ for an absolute constant C. The standardized third moment $\mathbb{E}|X_1 - m|^3/\sigma^3$ is bounded whenever σ is bounded away from 0; hence the approximation is $O(k^{-1/2})$ *uniformly over any family of problems with* $\sigma \geq \sigma_0 > 0$, which gives the one-variable collapse. The bound is *not* uniform in the degenerate limit $\sigma \to 0$ (both $p, w \to 0$), where F_k is anyway determined by the sign of m (Corollary F.3); we state the collapse for the non-degenerate regime and verify its empirical validity, including the residual's $k^{-1/2}$ decay, in §P.2. Ties $S_k = 0$ occur with probability $O(k^{-1/2})$ and are absorbed into R_k. □

K.4 Proposition F.9 (asymptotic gap identity)

Coverage. $1 - \mathrm{Cov}_k = \mathbb{E}[(1-p)^k]$; the integrand is bounded by 1 and converges pointwise to $\mathbf{1}\{p = 0\}$, so by bounded convergence $\mathrm{Cov}_k \to \mathbb{P}[p > 0]$.

Selection. Condition on the problem (a^*, θ) and apply the pointwise limits. If $m > 0$ then $p > (1-p)w$, i.e. the truth's mass strictly exceeds every wrong mass, and $\mathbb{P}[\text{plurality} = a^* \mid \theta] \to 1$ (Corollary F.3(i)). If $m < 0$ some wrong answer's mass strictly exceeds p, so the plurality is eventually among the maximal-mass wrong answers and $\mathbb{P}[\text{plurality} = a^* \mid \theta] \to 0$ (Corollary F.3(ii)); this requires no uniqueness of the wrong maximizer. If $m = 0$ and $p > 0$: by the uniqueness assumption exactly one wrong answer W ties the truth's mass, all other masses are strictly smaller, so eventually the plurality contest is between a^* and W; $N_{a^*} - N_W$ is a centered random walk with i.i.d. bounded non-degenerate steps (variance $p + (1-p)w > 0$), so $\mathbb{P}[N_{a^*} > N_W] \to \frac{1}{2}$ and $\mathbb{P}[N_{a^*} = N_W] = O(k^{-1/2}) \to 0$ (ties contribute 0 in the conservative convention, or their $\to 0$ mass under any tie-breaking), giving $\mathbb{P}[\text{plurality} = a^* \mid \theta] \to \frac{1}{2}$. The case $m = 0,\ p = 0$ forces $w = 0$, excluded by non-degeneracy. All limits are bounded by 1, so bounded convergence gives $F_k \to \mathbb{P}[m > 0] + \frac{1}{2}\mathbb{P}[m = 0, p > 0]$. Since $m > 0$ implies $p > 0$, subtracting yields $\lim(\mathrm{Cov}_k - F_k) = \mathbb{P}[p > 0] - \mathbb{P}[p > 0, m > 0] - \frac{1}{2}\mathbb{P}[p > 0, m = 0] = \mathbb{P}[p > 0, m < 0] + \frac{1}{2}\mathbb{P}[p > 0, m = 0]$. □

K.5 Theorem B.2 (rank zero–one law)

Fix the problem; write N_a for the count of answer a after k samples and $a_{(j)}$ for the wrong answer with the j-th largest share $q_{(j)}$. For a contest between a^* and a wrong answer a, the per-sample margin is $p - (1-p)q_a$, and Hoeffding (as in the proof of Corollary F.3) gives $\mathbb{P}[N_a \geq N_{a^*}] \leq e^{-k(p-(1-p)q_a)^2/2}$ when the margin is positive, and symmetrically $\mathbb{P}[N_{a^*} \geq N_a] \leq e^{-k(p-(1-p)q_a)^2/2}$ when it is negative.

(i) $m_L > 0$. The event $\{a^* \notin \text{top-}L\}$ requires at least L wrong answers to out-count a^*. Only $L-1$ wrong answers can have share exceeding $q_{(L)}$, so at least one out-counting wrong answer a has $q_a \leq q_{(L)}$, whence its contest margin is $p - (1-p)q_a \geq m_L > 0$. A union bound over the r wrong answers gives $\mathbb{P}[a^* \notin \text{top-}L] \leq r\, e^{-km_L^2/2}$.

(ii) $m_L < 0$. Then the L wrong answers $a_{(1)}, \ldots, a_{(L)}$ all have mass exceeding p. If every one of them out-counts a^*, then a^* is not in the top L; contrapositively, $\{a^* \in \text{top-}L\}$ requires $N_{a^*} \geq N_{a_{(j)}}$ for at least one $j \leq L$. Each such contest has margin $p - (1-p)q_{(j)} = m_j \leq m_L < 0$, so $\mathbb{P}[N_{a^*} \geq N_{a_{(j)}}] \leq e^{-km_j^2/2} \leq e^{-km_L^2/2}$; a union bound over $j \leq L$ gives the claim.

Limit. If no m_L vanishes, then $L \geq R^* \Leftrightarrow m_L > 0$ (the margins are nondecreasing in L and change sign at R^*), so (i) and (ii) give $\mathbb{P}[a^* \in \text{top-}L] \to \mathbf{1}\{L \geq R^*\}$, with both convergences exponential. Averaging over the problem prior gives the rank-CDF statement on the covered set (problems with $p = 0$ are never in any top-L for the truth). Necessity for Theorem F.7: $p < (1-p)w_2$ is $m_2 < 0$, and (ii) with $L = 2$ forces the top-two rate to 0. □

Exact exponents. The Hoeffding bounds above are not tight, and Sanov's theorem supplies the exact rate. The empirical measure $\hat{\nu} = N/k$ of an i.i.d. sample from θ on a finite alphabet satisfies a large-deviation principle with rate function $\mathrm{KL}(\cdot \,\|\, \theta)$. The event $\{a^* \notin \text{top-}L\}$ is $\bigcup_{|S|=L,\, a^* \notin S} \mathcal{C}_S$ with $\mathcal{C}_S = \{\nu : \nu_a \geq \nu_{a^*}\ \forall a \in S\}$, a finite union of closed convex sets. Each $\mathcal{C}_S$ is the closure of its interior, so its upper and lower Sanov bounds coincide and its rate is $\inf_{\mathcal{C}_S} \mathrm{KL}(\cdot\|\theta)$; a finite union

has the minimum of the rates, giving (3). The union bound is therefore tight to first order in the exponent, which is what licenses replacing "$\leq$" by "$= e^{-kI_L(1+o(1))}$".

For $L = 1$ and $S = \{b\}$ the program is solved in closed form. Constraints already slack at θ (i.e. $\theta_b \geq \theta_{a^*}$) give rate 0, so assume $\theta_b < \theta_{a^*}$; then the single constraint binds at the optimum, and writing $\nu_{a^*} = \nu_b = t$ and letting the remaining coordinates take their conditionally-proportional values, the objective is

$$h(t) \;=\; 2t \log \frac{t}{g} + (1-2t) \log \frac{1-2t}{1-\theta_{a^*}-\theta_b}, \qquad g = \sqrt{\theta_{a^*}\theta_b}.$$

Setting $h'(t) = 2\log\big[t(1-\theta_{a^*}-\theta_b)/(g(1-2t))\big] = 0$ gives $t^* = g/Z$ with $Z = 1-\theta_{a^*}-\theta_b+2g$, whence $1-2t^* = (1-\theta_{a^*}-\theta_b)/Z$ and

$$h(t^*) \;=\; \log \frac{1}{Z} \cdot \frac{2g+1-\theta_{a^*}-\theta_b}{Z} \;=\; -\log Z \;=\; -\log\big(1-(\sqrt{\theta_{a^*}}-\sqrt{\theta_b})^2\big),$$

which is (4). That this dominates $\frac{1}{2}(\theta_{a^*}-\theta_b)^2$ follows from $-\log(1-x) \geq x$ with $x = (\sqrt{\theta_{a^*}} - \sqrt{\theta_b})^2$ and $(\sqrt{\theta_{a^*}} - \sqrt{\theta_b})^2 \geq \frac{1}{2}(\theta_{a^*}-\theta_b)^2$ on the simplex. Taking the minimum over b gives I_1. □

Bahadur–Rao prefactor (Corollary F.3(ii)). In the two-mode reduction let $X_i \in \{-1, 0, 1\}$ with $\mathbb{P}[1] = p$, $\mathbb{P}[-1] = q$, $m = p - q < 0$, and $S_k = \sum_i X_i$. The moment generating function $M(\lambda) = pe^{\lambda} + qe^{-\lambda} + (1-p-q)$ is minimized at $e^{\lambda^*} = \sqrt{q/p}$, where $M(\lambda^*) = 1 - (\sqrt{p} - \sqrt{q})^2 = e^{-I(p,q)}$, confirming that the Chernoff exponent of $\mathbb{P}[S_k \geq 0]$ is $I(p, q)$. Under the tilted law $P^*(x) \propto e^{\lambda^* x} P(x)$ the walk is centred, with $P^*(\pm 1) = \sqrt{pq}/M(\lambda^*)$ and variance $\sigma_*^2 = 2\sqrt{pq}/M(\lambda^*)$. Since the walk is supported on the span-1 lattice, the Bahadur–Rao theorem in its lattice form gives

$$\mathbb{P}[S_k \geq 0] \;=\; \frac{M(\lambda^*)^k}{(1-e^{-\lambda^*})\,\sigma_* \sqrt{2\pi k}}\,(1+o(1)),$$

the factor $(1-e^{-\lambda^*})^{-1}$ arising from summing the geometric lattice weights $e^{-\lambda^* j}$, $j \geq 0$, rather than integrating a density. Substituting $e^{-\lambda^*} = \sqrt{p/q}$ gives the stated $C(p, q)$. □

L Proofs for Section 2

L.1 Theorem F.4

Let τ be the string-swap that exchanges the truth's and the wrong mode's identities. By construction the two modes have equal emission probability ($m = 0$) and the joint law of the full sample set (traces included) is invariant under τ composed with relabeling which mode is "correct." Hence for any feature Φ and any value ϕ, the posterior satisfies $\mathbb{P}[a^* = u \mid \Phi = \phi] = \mathbb{P}[a^* = v \mid \Phi = \phi]$ on the two-mode support, so $\max_a \mathbb{P}[a^* = a \mid \Phi] = 1/2$ there. By Theorem B.1, $T(\Phi) = \mathbb{E}[\max_{a \in \mathsf{A}(\Phi)} \mathbb{P}[a^* = a \mid \Phi]] = 1/2$ for every Φ, count- or trace-based, and every k. The true mode is absent only when all k draws emit the wrong mode, so $\mathrm{Cov}_k = 1 - 2^{-k} \uparrow 1$. □

L.2 Theorem F.5 (trace separation)

Write u, v for the two live modes and condition on the count vector. On the balanced stratum the counts have the same law under both truth-labelings (Theorem F.4), so they contribute nothing to the posterior and the problem reduces to a binary hypothesis test on the trace statistics with a uniform prior:

$$H_u : \; (t_i)_{i \in u} \sim P_c^{\otimes n_u},\; (t_i)_{i \in v} \sim P_w^{\otimes n_v}, \qquad H_v : \; (t_i)_{i \in u} \sim P_w^{\otimes n_u},\; (t_i)_{i \in v} \sim P_c^{\otimes n_v}.$$

By Theorem B.1 the ceiling is the Bayes accuracy of this test, which for a uniform binary prior is $\frac{1}{2}(1 + \mathrm{TV}(Q_u, Q_v))$ with Q_u, Q_v the two product laws (Tsybakov, 2009, Ch. 2).

The affinity factorizes. Because BC is multiplicative over product measures and symmetric in its arguments,

$$\mathrm{BC}(Q_u, Q_v) = \mathrm{BC}\big(P_c^{\otimes n_u}, P_w^{\otimes n_u}\big) \cdot \mathrm{BC}\big(P_w^{\otimes n_v}, P_c^{\otimes n_v}\big) = \rho^{\,n_u}\, \rho^{\,n_v} = \rho^{\,k'}.$$

Bounds. Le Cam's inequalities relate total variation to the affinity: $1 - \mathrm{BC} \le \mathrm{TV} \le \sqrt{1 - \mathrm{BC}^2}$. Substituting $\mathrm{BC} = \rho^{k'}$ and using $T = \frac{1}{2}(1 + \mathrm{TV})$ gives

$$1 - \tfrac{1}{2}\rho^{k'} \;\le\; T(\Phi_{\mathrm{tr}}) \;\le\; \tfrac{1}{2}\big(1 + \sqrt{1 - \rho^{2k'}}\big),$$

as claimed. (The lower bound is vacuous, i.e. below $\frac{1}{2}$, when $\rho^{k'} > 1$, which cannot occur; it is weaker than the trivial bound $T \ge \frac{1}{2}$ only when $\rho^{k'}$ is near 1.)

(i) Dichotomy. If $\rho = 1$ then $\mathrm{BC}(P_c, P_w) = 1$, which for probability measures forces $P_c = P_w$, hence $Q_u = Q_v$, hence $\mathrm{TV} = 0$ and $T = \frac{1}{2}$ for every k'. Conversely if $\rho < 1$ the upper bound is < 1 only for finite k' while the lower bound $1 - \frac{1}{2}\rho^{k'}$ exceeds $\frac{1}{2}$ as soon as $\rho^{k'} < 1$, so the ceiling is strictly above chance.

(ii) Rate. $\rho^{k'} = e^{-k' \log(1/\rho)}$, and the lower bound gives $1 - T \le \frac{1}{2} e^{-k' \log(1/\rho)}$.

(iii) Additivity. Off the balanced stratum the counts are informative. Bayes' rule on the joint feature gives, with $\lambda = \log dP_c/dP_w$,

$$\log \frac{\mathbb{P}[a^* = u \mid \Phi_{\mathrm{tr}}]}{\mathbb{P}[a^* = v \mid \Phi_{\mathrm{tr}}]} = \underbrace{\log \frac{\mathbb{P}[a^* = u \mid \Phi_{\mathrm{ct}}]}{\mathbb{P}[a^* = v \mid \Phi_{\mathrm{ct}}]}}_{\Lambda(\Phi_{\mathrm{ct}})} + \sum_{i \in u} \lambda(t_i) \;-\; \sum_{i \in v} \lambda(t_i),$$

since the trace statistics are conditionally independent of the counts given the labeling. Under H_u the summands have means $\mathbb{E}_{P_c}[\lambda] = D(P_c \| P_w)$ and $-\mathbb{E}_{P_w}[\lambda] = D(P_w \| P_c)$ respectively, both nonnegative by Gibbs' inequality, giving the stated drift. □

L.3 Theorem F.7

Top-two. Order the emission probabilities $\theta_{(1)} \ge \theta_{(2)} \ge \cdots$. If $p > (1-p)w_2$ then θ_{a^*} exceeds every wrong mass except possibly the largest, so $\{\theta_{a^*}\}$ is among the top two values; by the SLLN the empirical top-2 counts coincide with the top-2 masses a.s. for large k, hence $\mathbb{P}[a^* \in \text{top-2 counts} \mid \text{covered}] \to 1$, i.e. $S(2) \to \mathrm{Cov}_k$. On the symmetric stratum both modes are top-2 by construction, so $S(2) = \mathrm{Cov}_k = 1 - 2^{-k}$ while $S(1) = 1/2$ by Theorem F.4. □

L.4 Lemma F.6 (pooling inflates the affinity)

The function $(x, y) \mapsto \sqrt{xy}$ is concave on $\mathbb{R}^2_{\ge 0}$: its Hessian is $-\frac{1}{4}(xy)^{-3/2}\begin{pmatrix} y^2 & -xy \\ -xy & x^2 \end{pmatrix}$, negative semidefinite. Hence for each point t of the common dominating measure, $\sqrt{(\sum_j \pi_j P_c^{(j)}(t))(\sum_j \pi_j P_w^{(j)}(t))} \ge \sum_j \pi_j \sqrt{P_c^{(j)}(t) P_w^{(j)}(t)}$ by Jensen; integrating over t gives the claim, since BC is the integral of that pointwise quantity. Equality holds iff the ratio $P_c^{(j)}/P_w^{(j)}$ is the same for all j with $\pi_j > 0$, i.e. exactly under the homogeneity that Theorem F.5 assumes. Consequently a pooled estimator targets BC of the mixtures, which is at least the mean per-problem affinity, so heterogeneity moves $\hat{\rho}$ toward 1 and the affinity-based ceiling bounds toward $\frac{1}{2}$. □

M Reproducibility

Analysis implementation. All estimators are implemented in a single shared library: ceilings (including rank rules and list ceilings), error concentration, margin and dominant mass, the exact large-deviation rates and the sharp trace exponent, and the synthetic generators. The implementation is covered by an automated test suite (254 tests, all passing). Every run is seeded, every quantitative figure and table is rendered from the recorded experimental outputs rather than transcribed by hand, and the quantities typed in the prose were cross-checked against those outputs.

Theory, verified independently of its proofs. The core statements were checked numerically, without reference to their own proofs. The ceiling identity $T = 2^{-H_\infty}$ and the gap identity (Proposition F.9) were confirmed to machine precision, and concavity of G_L exactly; the rank zero–one law and the margin CLT against Monte Carlo (≤ 0.02); the verification value of Theorem B.1(iii) against

exhaustive enumeration of all B-bit protocols ($B \leq 2$) to 0 error; and the trace-separation theorem (T7) by enumerating the whole outcome space of a two-mode trace experiment, confirming that the affinity factorizes as $\rho^{k'}$ ($< 10^{-14}$), that the exact Bayes accuracy lies inside the Le Cam bounds (0 violations), and that a mode-symmetric trace law pins the ceiling at exactly $1/2$ ($< 10^{-15}$). A second set of checks covers the exact Sanov exponents against a simplex search and the closed-form pairwise Chernoff value (V3), the lattice Bahadur–Rao prefactor against exact convolution (V4), the sharp trace exponent (Theorem F.5(iv)) with its balanced/unbalanced dichotomy (V8), and the pooling inequality (Lemma F.6) on random mixtures (V9).

Experiments. The synthetic results are generated at fixed seeds. The four full-n cells of §3 are sampled once per cell and then analysed in stages (first-order sampling, the meta and judge elicitations, decomposition, count ceiling, Step-3 selectors, trace separation); each cell records its seed, temperature, k, the locked sampling configuration, the serving endpoint and the outcome of the completeness check. Derived quantities (the headline table, the k-sweep, the guessing curve, the difficulty strata, the rank distribution and the Step-3 power ledger) are computed from those samples, as is the extraction-sensitivity analysis, which also re-verifies on every run that no generation reached the token cap. The pool analyses are computed the same way; the trace-separation analysis runs on the own-sampled pilot cell of §P.7, which is reported separately from the four grid cells and never pooled with them.

Gates. Three gates run ahead of every grid analysis and are recorded per cell. *Completeness*: no analysis stage may start unless n, k and problem-id uniqueness match the specification; duplicate ids, missing log-probs and corrupt records are reported. *Oracle*: the analysis aborts rather than warns unless the answer clustering reproduces the generation-time `math_verify` score on every sample; it reads 1.000000 on all four cells and caught two grading defects during this run, both logged with their fixes. *Token cap*: zero of 232,832 generations reached the 1024-token cap, the longest being 740 tokens. Label-freeness is structural rather than reviewed: the selector's view and the gold label are constructed as separate objects and gold is asserted to be absent from the view, and a test verifies that permuting which answer is gold changes no selector's choice. The protocol, the predictions and the stop condition were fixed in a written pre-registration before any real-data outcome was observed; every deviation, post-hoc addition and failed prediction is recorded there with a date and a reason.

N SCOPE CONDITIONS

The body's limitations paragraph states the two boundaries that change what the result means. The remaining conditions bound how precisely it is measured, and are collected here.

Sample sizes. The budget axis and the cross-family breadth come from the pool cells, which carry k to 10^4 but only $n \approx 128$ problems, so their bootstrap intervals are wide; the benchmark-scale problem counts come from the grid cells, which stop at k=64. The two families are reported separately throughout and never pooled, so no interval in the paper mixes the two error scales.

Selector scope. The label-free negative of §6 covers the count-profile family in full (a cross-fitted gradient-boosted rule on all 10 cells), the seven pre-registered trace-aware selectors on the four grid cells, and the feature-level affinity of three per-sample trace statistics. Within that family we find no exploitable rank miscalibration at these sample sizes, with stated power on the grid cells. A feature outside it, such as a learned per-sample scorer, a process-level signal, or a second model's traces, is not bounded by these measurements, only by Theorem 2.1, which permits it to reach the trace-level ceiling and prices what that would be worth.

Verifier scope. The verification measurements of §7 run up to 7B parameters, on two benchmarks, in one model family. They establish that a model's own verification pass does not escape its own error structure at that scale; they do not establish that no verifier can. A stronger or differently-correlated verifier is exactly the external information Corollary 2.5 prices, and the right experiment is to measure how many of its B bits are noiseless, not to assume the bound is unreachable.

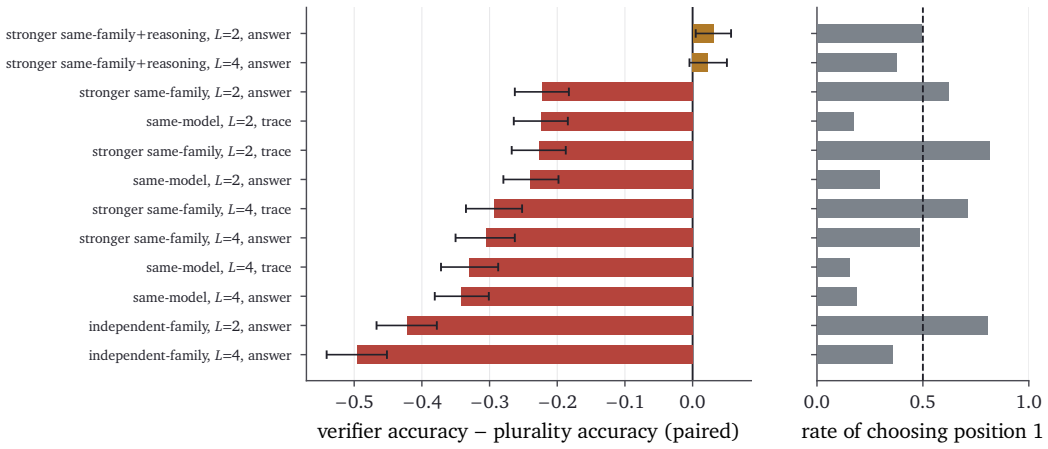


Figure 6: **Every implementable verifier against the same count-only baseline, on the same problems.** GSM8K/Qwen2.5-3B, $k = 16$, multi-candidate slice ($n = 651$), paired bootstrap. *Left:* only the generative condition (gold) improves on counting; every index-only condition (red) loses 22 to 50 points. *Right:* the mechanism. Candidate order is randomized, so a content-driven selector at $L{=}2$ would sit at the dashed line. The index-only verifiers do not, and the generative one does.

Causal evidence. The mechanism is causal where we can intervene synthetically (Appendix P: the synthetic generators move R^* by construction and the frontier follows) and predictive on real models (§4). An intervention that moves R^* on a fixed model and a fixed benchmark, by decoding temperature, prompt, or a targeted fine-tune, is what closes the remaining inferential gap. Appendix P.10 reports one: with the checkpoint and the decoding held fixed, toggling Qwen3-8B's thinking mode moves the truth-rank distribution, and the independently measured selection frontier moves with it, consistent with the rank law under its calibrated null. Being a single checkpoint on a single task, it narrows this gap rather than closing it.

O EXPERIMENT B IN FULL: WHAT AN IMPLEMENTABLE VERIFIER ACTUALLY BUYS

This appendix is the full version of §7: the complete condition list, the preregistered secondary tests, the adaptive controller, and the sensitivity checks. Section 2 prices identification in *capacity*: a target-aware transcript with $M{=}2^B$ attains exactly G_M, and the count-ranked shortlist realization used here attains $S_M(k) \le G_M$, with equality under Corollary 2.4. Capacity is an upper bound on an idealized object, and the gap between it and a deployable system is the whole practical question. This section closes that gap empirically. Two layers are kept strictly apart throughout. **Layer A** uses gold labels to measure ceilings (the count-ranked shortlist mass S_L, which lower-bounds G_L; the truth's count rank; coverage). **Layer B** runs real verifiers that never see gold: each receives the problem, a count-ranked shortlist of L candidate answers in a deterministically randomized order, optionally one representative trace per candidate, and must return a candidate index. A model call is never called "a bit."

O.1 THE HEADLINE: CAPACITY IS REAL, AND CHEAP VERIFICATION CANNOT REACH IT

Table 5 reports every verifier condition against the *same* count-only baseline on the *same* problems, paired at the problem level. We report the multi-candidate slice, problems whose observed shortlist holds at least two distinct candidates. Single-candidate problems are those on which a verifier has no decision to make; including them dilutes every contrast toward zero and flatters the verifier.

The slice matters and we state its size: at $k = 16$ on GSM8K/Qwen2.5-3B, 668 of 1,319 problems produce a single distinct answer, on which plurality is already right 98.7% of the time and a verifier has nothing to decide. The 651 remaining problems are where identification actually happens, and they are where the coverage–selection gap lives.

Table 5: **Every implementable verifier, against the same count-only baseline on the same problems.** Multi-candidate slice, $k = 16$, paired problem-level bootstrap (10,000 resamples). G_L is the oracle probability that the truth is somewhere in the shortlist – the headroom a perfect selector could take. "pos. 1" is the rate at which the verifier picks the first listed candidate under randomized order; at $L = 2$, content-driven selection would give 0.50. Of the conditions listed, exactly one improves on counting: the generative verifier on GSM8K/3B, where the verifier outranks the sampler on a benchmark it can solve. Every index-only condition loses, and so does every generative condition in which the verifier is the sampler's own model.

Verifier	Cell	L	prompt	n	plur.	G_L	verifier	Δ vs plurality	pos. 1
same-model (3B)	GSM8K / 3B	2	answer	651	0.803	0.905	0.564	-0.240 [-0.280, -0.198]	0.29
same-model (3B)	GSM8K / 3B	2	trace	651	0.803	0.905	0.579	-0.224 [-0.264, -0.184]	0.17
same-model (3B)	GSM8K / 3B	4	answer	651	0.803	0.925	0.461	-0.343 [-0.381, -0.301]	0.19
same-model (3B)	GSM8K / 3B	4	trace	651	0.803	0.925	0.473	-0.330 [-0.372, -0.287]	0.16
same-model (3B)	MATH / 3B	2	answer	370	0.630	0.711	0.416	-0.214 [-0.265, -0.162]	0.37
same-model (3B)	MATH / 3B	2	trace	370	0.630	0.711	0.419	-0.211 [-0.259, -0.162]	0.25
same-model (3B)	MATH / 3B	4	answer	370	0.630	0.768	0.330	-0.300 [-0.354, -0.243]	0.17
same-model (3B)	MATH / 3B	4	trace	370	0.630	0.768	0.311	-0.319 [-0.376, -0.262]	0.12
stronger, same family (7B)	GSM8K / 3B	2	answer	651	0.803	0.905	0.581	-0.223 [-0.263, -0.183]	0.62
stronger, same family (7B)	GSM8K / 3B	2	trace	651	0.803	0.905	0.576	-0.227 [-0.267, -0.187]	0.81
stronger, same family (7B)	GSM8K / 3B	4	answer	651	0.803	0.925	0.498	-0.306 [-0.350, -0.263]	0.48
stronger, same family (7B)	GSM8K / 3B	4	trace	651	0.803	0.925	0.510	-0.293 [-0.335, -0.252]	0.71
stronger, same family (7B)	MATH / 3B	2	answer	370	0.630	0.711	0.484	-0.146 [-0.195, -0.100]	0.56
stronger, same family (7B)	MATH / 3B	2	trace	370	0.630	0.711	0.414	-0.216 [-0.270, -0.165]	0.71
stronger, same family (7B)	MATH / 3B	4	answer	370	0.630	0.768	0.370	-0.259 [-0.314, -0.205]	0.41
stronger, same family (7B)	MATH / 3B	4	trace	370	0.630	0.768	0.346	-0.284 [-0.341, -0.224]	0.51
independent family (3B, 4-bit)	GSM8K / 3B	2	answer	651	0.803	0.905	0.381	-0.422 [-0.467, -0.378]	0.80
independent family (3B, 4-bit)	GSM8K / 3B	4	answer	651	0.803	0.925	0.307	-0.496 [-0.541, -0.452]	0.36
generative (7B, reasoning)	GSM8K / 3B	2	answer	651	0.803	0.905	0.834	+0.031 [+0.005, +0.057]	0.49
generative (7B, reasoning)	GSM8K / 3B	4	answer	651	0.803	0.925	0.826	+0.023 [-0.005, +0.051]	0.38
generative (7B, reasoning)	GSM8K / 7B	2	answer	375	0.784	0.909	0.675	-0.109 [-0.155, -0.064]	0.50
generative (7B, reasoning)	GSM8K / 7B	4	answer	375	0.784	0.939	0.632	-0.152 [-0.200, -0.104]	0.46
generative (7B, reasoning)	MATH / 3B	2	answer	370	0.630	0.711	0.516	-0.114 [-0.162, -0.068]	0.41
generative (7B, reasoning)	MATH / 3B	4	answer	370	0.630	0.768	0.459	-0.170 [-0.219, -0.122]	0.33
generative (7B, reasoning)	MATH / 7B	2	answer	307	0.606	0.746	0.453	-0.153 [-0.215, -0.094]	0.43
generative (7B, reasoning)	MATH / 7B	4	answer	307	0.606	0.805	0.381	-0.225 [-0.287, -0.163]	0.31

The oracle headroom is real: on this slice plurality reaches 0.803 while the truth lies somewhere in the top-2 shortlist 0.905 of the time. A perfect one-of-two selector would gain 10.2 points.

Every *index-only* verifier instead *loses* between 22 and 50 points. This is not a small shortfall; it is a reversal. The mechanism is visible in the last column: with candidate order randomized by construction, these verifiers select position 1 between 16% and 81% of the time rather than the 50% (at L=2) that content-driven selection would produce. They are not reading the candidates so much as responding to their position, and in doing so they overwrite a count prior that was already right four times in five. Verifier capability alone does not fix this: replacing the 3B verifier with a 7B one from the same family moves the multi-candidate accuracy from 0.564 to 0.581 (L=2, answers only), leaving it 22 points below simply counting.

O.2 Reasoning changes the sign

The index-only protocol forbids the verifier from re-deriving the answer, which is precisely what generative verifiers do. We therefore preregistered a *generative* condition (prediction P-B9): the same 7B verifier is told to work the problem through and end with a final index line, with a deliberately lenient parser that favours the verifier. Parse rate is 1.000 and replies average 233.9 output tokens.

On GSM8K this condition is the only one that beats counting. On the multi-candidate slice it reaches 0.834 against plurality's 0.803, a paired +0.031 with 95% CI [+0.005, +0.057]; over the full cell it is +0.015 [+0.002, +0.028]. Its position-1 rate is 0.49, and the position bias disappears exactly when the model actually computes. It realizes about 30% of the available oracle headroom.

The second clause of P-B9 was therefore falsified on GSM8K, and as the preregistration required we report that falsification as the outcome rather than retaining the negative headline.

Table 6: **The generative verifier is a lossy channel for its own answer posterior.** Every `genverify` cell: the verifier is Qwen2.5-7B-Instruct and the same model's $k = 64$ bank supplies its own answer posterior q restricted to the shortlist. Multi-candidate, parse-valid, truth-in-shortlist slice; paired problem-level bootstrap (10,000 resamples). ΔLL is the mean log-likelihood a conditional logit loses when that regressor is dropped: removing ground-truth correctness costs almost nothing. "follow q" reads the shortlist candidate the verifier's own samples most support instead of asking it to judge, and beats the judge in every cell; "k=2" does the same with only two fresh samples, at 0.71–0.88× the judge call's tokens.

		ΔLL if dropped		accuracy		Δ vs judge	
Cell	n	correct	log q	judge	follow q	follow q	k=2
GSM8K / 3B / L2	589	0.0028	0.0304	0.922	0.954	+0.032 [+0.012, +0.054]	+0.003 [-0.016, +0.024]
GSM8K / 3B / L4	602	0.0001	0.0678	0.894	0.939	+0.045 [+0.022, +0.068]	+0.011 [-0.009, +0.032]
GSM8K / 7B / L2	341	0.0000	0.0167	0.742	0.845	+0.103 [+0.053, +0.152]	+0.047 [+0.004, +0.089]
GSM8K / 7B / L4	352	0.0003	0.0349	0.673	0.821	+0.148 [+0.099, +0.199]	+0.075 [+0.033, +0.118]
MATH / 3B / L2	263	0.0000	0.0237	0.726	0.943	+0.217 [+0.156, +0.278]	+0.152 [+0.099, +0.208]
MATH / 3B / L4	283	0.0028	0.0341	0.601	0.890	+0.290 [+0.230, +0.346]	+0.210 [+0.159, +0.261]
MATH / 7B / L2	229	0.0070	0.0052	0.607	0.847	+0.240 [+0.162, +0.319]	+0.140 [+0.072, +0.208]
MATH / 7B / L4	247	0.0027	0.0027	0.474	0.789	+0.316 [+0.243, +0.389]	+0.186 [+0.124, +0.248]

But it does not transfer to the harder benchmark, and that is the most informative result in this section. On MATH-500 the same generative verifier reaches only 0.516 against plurality's 0.630, a paired -0.114 $[-0.162, -0.065]$ at L=2 and -0.170 $[-0.219, -0.122]$ at L=4. It is the best verifier we ran on that cell (better than every index-only condition, which span -0.146 to -0.319), and it is still far below simply counting.

And it disappears entirely when the verifier is the sampler. The GSM8K/3B condition pairs a 3B sampler with a 7B verifier. Running the same generative protocol where the verifier *is* the sampling model (Qwen2.5-7B judging its own samples), gives 0.675 against plurality's 0.784 on GSM8K, a paired -0.109 $[-0.155, -0.064]$, and 0.453 against 0.606 on MATH-500, -0.153 $[-0.215, -0.094]$. The GSM8K position-1 rate is exactly 0.50: this verifier is reasoning, unbiased, and still worse than counting.

Figure 5 (left) shows all eight generative conditions. **One is positive**, and it is the one whose verifier outranks its sampler on the benchmark the verifier can solve. Its L=4 companion is the only other non-negative cell and its interval contains zero. The remaining six lose between 10.9 and 22.5 points.

Taken together the conditions identify the mechanism, and it is the paper's own. Verification helped in exactly one configuration, and that configuration had a verifier *stronger than the sampler* on a task the verifier could solve. Remove either property, either by making the task hard (MATH-500) or by making the verifier the same model as the sampler (both 7B cells), and the gain becomes a loss.

The same-model result is the one that matters, and it is not a story about prompt format, shortlist depth, trace access, or position bias, all of which we varied or measured away. It is the correlated-error mechanism of §2 reappearing on the selection side. A model that voted for a wrong answer k times does not stop believing it when asked to check; the verification pass is drawn from the same error structure that produced the plurality, so it re-derives the same mistake and then overrides a count prior that at least aggregated over k independent draws. Self-verification is not an external information source. It is another sample, priced like a long one.

O.3 What the generative verifier is actually reading

The paragraph above is an interpretation, and the data already collected are enough to test it. In every generative condition the verifier is Qwen2.5-7B-Instruct, and the same model was sampled $k = 64$ as a *generator* on the same problems. For each problem we can therefore form the verifier's own answer posterior restricted to the shortlist it was shown, $q(c \mid x) \propto n_c + \frac{1}{2}$, where n_c counts its own samples landing in candidate cluster c, and ask what its selection is a function of. In the 7B-sampler cells the shortlist was built from the first 16 samples of that same bank, so q is read off the held-out remainder. Throughout this subsection the slice is multi-candidate, parse-valid, and truth-in-shortlist, which is where a selector both has a decision to make and can be right by making it. The join is

validated rather than assumed: recomputing verifier correctness from the frozen banks reproduces the per-problem records on all 7,276 rows.

Ground-truth correctness is screened off by the verifier's own posterior. In a conditional logit over shortlist alternatives containing $\log q$, the generator's log count, position, and trace length, adding an indicator for which candidate is actually correct buys at most 0.0070 mean log-likelihood in any cell and 0.0001 pooled, while dropping $\log q$ costs 0.0449 pooled. Conditional on what the model would have answered, whether a candidate is right is close to irrelevant to which candidate it picks.

Its departures from its own posterior are worse than useless. Pooled over $2{,}906$ problems the judge departs from $\arg\max_S q$ on 680 of them (23.4%). On exactly those problems it is right 0.179 $[0.151, 0.210]$ of the time, against 0.775 $[0.743, 0.806]$ had it simply followed q. The departure rate tracks the posterior margin (0.62, 0.52, 0.50, 0.47 across margin bins below 0.65, and 0.14 above it), which is what a noisy draw from q predicts and what deliberation over candidate content does not.

So reading the verifier beats asking it, at lower cost. Replacing the judge call with "keep the shortlist candidate the verifier's own samples most support" wins in 8 of 8 cells, by $+0.032$ to $+0.316$, every interval excluding zero (Table 6). Two fresh samples suffice: at $k = 2$ the rule wins in all eight cells on the point estimate and in six with intervals excluding zero, for 0.71 to $0.88\times$ the judge call's tokens. Ties are broken uniformly at random, using the verifier's own samples and nothing else; breaking them by the generator's count order instead raises the pooled gain from $+0.079$ $[+0.064, +0.093]$ to $+0.121$ $[+0.105, +0.136]$, and we report the weaker number as primary because the stronger one quietly restores the count prior the contrast is about.

This is the sharp form of the mechanism above. The generative verifier is not an adjudicator that happens to be poor; it is a lossy channel for its own answer distribution, and the loss is what the verification pass contributes. Two caveats bound the claim. First, q is measured at temperature 0.8 from the generation bank while the judge ran at temperature 0.0 with the shortlist in context, so q is not the judge's own decoding gauge; the $k = 2$ contrast, which compares two real sampling policies, is the gauge-robust version. Second, screening-off is only *identified* where q and the truth disagree. In the two MATH/7B cells, where the verifier is weakest, dropping correctness costs as much as dropping $\log q$ (0.0070 against 0.0052 at $L = 2$), and we do not claim screening-off there.

The bottom line, across the 26 implementable verifier conditions we ran (four verifier roles, two shortlist depths, two prompt forms, four model–task cells): *verification is never the better use of marginal inference compute.* In the single condition where it beat counting, the same budget spent on the stronger generator beat it in turn (Table 7); wherever the generator was already the strongest model available, verification lost outright.

The negative has a precise reach. It is not that verification cannot help: the oracle headroom is real, 10.2 points on the GSM8K/3B slice, and one condition collects about a third of it. It is that up to 7B, on two math benchmarks in one model family, a model's own verification pass does not escape its own error structure, and the one gain is explained by verifier–sampler capability asymmetry that the generation side buys more cheaply. *Cross-model* consensus (Liu, 2026) is the direction our mechanism predicts should work; the independent checkpoints available to us were gated or quantized.

Deeper is not better, in every cell. Going from $L{=}2$ to $L{=}4$ raises oracle capacity in all four cells and *lowers* realized accuracy in all four. On GSM8K/3B capacity rises $0.905 \to 0.925$ while the advantage over plurality weakens from $+0.031$ to $+0.023$ $[-0.005, +0.051]$; on MATH-500/3B the move costs 5.6 points ($-0.114 \to -0.170$) against capacity $0.711 \to 0.768$; on GSM8K/7B it costs 4.3 ($-0.109 \to -0.152$) against $0.909 \to 0.939$; on MATH-500/7B it costs 7.2 ($-0.153 \to -0.225$) against $0.746 \to 0.805$. Capacity and realization move in opposite directions with depth in all four cells: a longer shortlist offers the selector more truth to find and more ways to be wrong, and at these verifier scales the second effect wins. Useful verification depth is therefore shallow, and it is not predicted by the measured capacity, which is exactly the mechanism behind the failure of P-B6 below.

No bound in Section 2 predicts this, and we flag that plainly: every capacity statement there is monotone in L, because capacity is a property of the shortlist and not of any selector that must read it. The observed non-monotonicity is therefore not a violation of the theory but a measurement of the distance between the theory's object and a deployable one. Read together with the 30% realization

Table 7: **The control that scopes the prescription.** Sampling-only accuracy and measured output-token cost on GSM8K. Sample-16-then-generatively-verify (3B sampler, 7B verifier) reaches 0.913 for $3{,}120$ tokens: better than any 3B sampling budget at comparable cost, but dominated by simply sampling the 7B model eight times (0.918 for $1{,}154$). Within a fixed sampler, verification wins after the plateau; across samplers, a better generator wins.

	Qwen2.5-3B sampler		Qwen2.5-7B sampler	
Policy	accuracy	tokens	accuracy	tokens
self-consistency, $k = 1$	0.820	170	0.888	144
self-consistency, $k = 4$	0.864	680	0.904	576
self-consistency, $k = 8$	0.891	1,362	0.918	1,154
self-consistency, $k = 16$	0.898	2,720	0.928	2,308
self-consistency, $k = 32$	0.904	5,447	0.924	4,611
self-consistency, $k = 64$	0.908	10,891	0.925	9,226
$k = 16 \rightarrow$ verify (generative)	**0.913**	**3,120**	–	–
$k = 16 \rightarrow$ verify (generative)	–	–	**0.897**	–

figure, the empirical content of this section is that *transcript capacity is a very loose upper bound on implementable selection*, and this section measures how loose it is at these scales.

O.4 Matched real cost, and the control that decides the prescription

At $k = 16$ on GSM8K/Qwen2.5-3B, generation costs $2{,}720$ output tokens per problem and the generative verifier adds 166.4 input plus 233.9 output tokens, for $3{,}120$ total, equivalent to sampling only $k \approx 18$. Against the sampling-only frontier, sample-then-verify is $+0.015$ $[+0.002, +0.028]$ over k=16 at $1.15\times$ the tokens; $+0.009$ $[-0.004, +0.022]$ over k=32 at $0.57\times$; and $+0.005$ $[-0.008, +0.018]$ over k=64 at $0.29\times$. Read against the locked dominance rule, the load-bearing difference in the last two rows is *cost*: the pipeline matches the accuracy of 64-sample self-consistency for 29% of its tokens.

Read alone, that looks like a systems result. It is not, and two controls dissolve it.

Control 1: a better generator is cheaper than a better selector. Our verifier is larger than our sampler, so the comparison that decides the prescription is not only against more 3B samples but against spending the same budget on the 7B model directly. Sampling the 7B model eight times reaches 0.918 for $1{,}154$ tokens, while 3B-sample-then-7B-verify reaches 0.913 for $3{,}120$ (Table 7). The hybrid is *Pareto-dominated* by simply using the better generator, and remains so under a parameter-weighted proxy ($8{,}078$ versus $10{,}960$ billion-parameter-tokens; a proxy, not a FLOP measurement).

Control 2: the gain does not exist once you are already using your best model. Verification with the sampler itself loses 10.9 points on the same benchmark. So the $3{,}120$-token pipeline is not buying verification; it is buying access to a stronger model through an expensive channel.

The prescription therefore does not survive in the form the literature debates. *Nothing in our data supports moving inference compute from generation to verification*, and the surviving positive cell is explained by verifier–sampler capability asymmetry rather than by verification. This is consistent in direction with Singhi et al. (2025), who find self-consistency more compute-efficient than generative verification over most practical budgets, and with Liu (2026), who report that a model scoring its own candidates closes almost none of the gap while *cross-model* consensus closes much of it. Our same-model result is the sharp version of that observation, and our theory says why: decorrelation, not adjudication, is what buys identification.

O.5 Where the benefit lives, and where the diagnostics fail

Three preregistered secondary tests sharpen the picture and two of them fail.

P-B3 (confirmed). Verification benefit concentrates exactly where the theory says identification is hard: on problems whose truth is not count-rank 1, the benefit exceeds that on rank-1 problems

by $+0.44$ to $+0.55$ across cells and depths, every interval excluding zero. This is a label-stratified explanation, not a deployable rule, because the rank is not observable at test time.

P-B4 (not confirmed). The independent-family verifier does not disproportionately help shared-wrong-mode problems: the interaction is -0.001 at $L{=}2$ and $+0.122$ $[-0.020, +0.259]$ at $L{=}4$, so we do not claim verifier independence as the mechanism. The available independent model was a 4-bit quantized 3B surrogate, which bounds what this arm settles either way.

P-B6 (falsified). The measured count-ranked shortlist mass S_L does not predict useful verification depth better than benchmark identity: the differences in outer-fold log loss and Brier score are -5.4×10^{-4} and -3.6×10^{-5}, both favouring the benchmark dummy ($p \approx 0.98$). The appealing story that a measured curve tells a system how deep to verify is not supported by our data.

O.6 The adaptive controller wins by not sampling, not by verifying

The cross-fitted controller chooses among `ANSWER`, `SAMPLE_MORE`, `VERIFY_TOP2`, `VERIFY_TOP4` from observable count statistics only, is frozen before the held-out fold, and is evaluated on outer-fold predictions alone. On the two Qwen2.5-7B cells, 16 configurations meet the locked dominance rule: on GSM8K it saves $4{,}570$ tokens per problem (about half the $k{=}64$ budget) at an accuracy difference of $+0.008$ $[+0.002, +0.014]$; on MATH-500 it saves $6{,}930$ $[6{,}389, 7{,}468]$ tokens at $+0.006$ $[-0.012, +0.024]$. P-B5 is supported on those cells and not on the two 3B cells.

How it wins matters. Verification actions are rare in the frozen controllers; the savings come from recognizing, from the count profile alone, that further sampling will not change the answer, and stopping. That is a useful deployable rule, and it is a rule about generation, not about verification.

O.7 Sensitivity

Reversing the deterministic candidate order changes no conclusion. Parse failures remain in the denominator as errors under intention-to-treat; available-case accuracy is reported secondarily and is never load-bearing. Latency is recorded per condition but the independent-family arm ran on a different backend and precision, so no cross-backend latency frontier is reported; tokens are the primary cost endpoint and GPU-seconds and dollars remain NA rather than imputed.

P Extended experimental results

This appendix is the full version of §3: the synthetic validation of the mechanism (Steps 0 and 0b), the cross-family replication on the reused pool, the dataset-level count-only ceiling, the de-circularized margin law and the extractor-noise falsification test (Step 2), and the selector and trace-feature results at length. The step numbering follows the pre-registration rather than a narrative order. Every figure and table is generated from the recorded experimental outputs.

P.1 Step 0, synthetic validation of the mechanism

Before touching real models we verify that the mechanism, and our estimators of it, behave as the theory demands. We generate sample sets with a tunable error-generating process: each of k samples is correct with probability p; a wrong sample takes a designated shared wrong answer with probability ρ and is otherwise drawn from a large diffuse pool ($n{=}2000$ problems per cell, fixed seeds). This is the single-dominant family, on which the theory makes exact predictions.

Both predicted regimes appear exactly (Figures 7, 8). Sweeping ρ at fixed p collapses the frequency ceiling to 0.0005 while coverage holds at 1.00 and pass@1 is pinned by construction: the entire gap is manufactured by error correlation alone, and the Condorcet signature, the dominant wrong answer strictly out-voting the truth, tracks it to 99.95% of covered problems. Compute makes it worse, as Corollary F.3(ii) requires: at $\rho = 0.85$, growing k from 4 to 512 drives F from 0.25 to 0.00 while coverage rises to 1.00; at $\rho = 0.10$ the same sweep shows the success regime. The estimators recover the mechanism, which gates proceeding to real data.

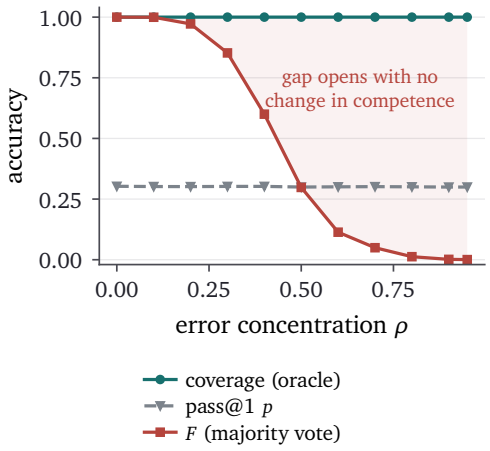


Figure 7: **Error concentration collapses selection while coverage is unmoved.** Synthetic sweep at fixed $p = 0.30$, $k = 64$, n=2000 problems per point. Majority-vote accuracy F falls from 1.00 at $\rho = 0$ to 0.0005 at $\rho = 0.95$; coverage stays at 1.00; pass@1 is flat at 0.30 by construction. The gap is created entirely by the correlation structure of errors, not by any change in model competence.

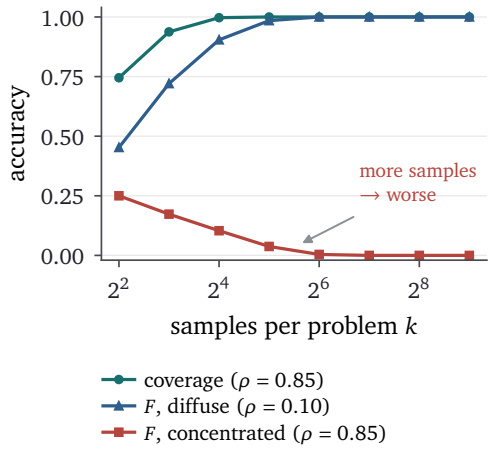


Figure 8: **More sampling does not rescue voting; it entrenches the loser.** Frequency ceiling vs. k at $p = 0.30$ With diffuse errors ($\rho = 0.10$), F rises to 1.00 by $k = 64$ (Condorcet success, Corollary F.3(i)). With concentrated errors ($\rho = 0.85$), F falls monotonically from 0.25 at $k = 4$ to 0.004 at $k = 64$ and 0.00 at $k \geq 128$, while coverage rises to 1.00, the exponential entrenchment of Corollary F.3(ii).

P.2 Step 0b, rank structure and the attained ceiling

The synthetic family above validates that voting *fails*; it does not, by itself, show that the information is *lost*, and it is not. Two further experiments (no LLM) separate the two, matching the revised theory.

Rank structure (Fig. 9). On the same family, as ρ rises the truth migrates from count-rank 1 to rank 2, so the label-free runner-up rule inverts the plurality's fate (0.9985 against 0.0015 at $\rho = 0.9$) while $S(2)$ holds at 1.00. The gap in this regime is count-recoverable: a decoding failure of plurality, not an information loss (Theorem B.1).

The attained ceiling (Fig. 10). Genuine information collapse needs the margin-zero family: truth and one wrong answer at *equal* emission probability. There every count-only rule scores 0.51 ± 0.01, the Le Cam $1/2$ attained, confirming Theorem F.4, while $S(2) \geq 0.98$ confirms Theorem F.7 and the one-bit closure of Proposition F.8. This is the family the impossibility claim requires, and where a trace feature breaking the truth/wrong-mode symmetry is the *only* label-free escape.

The parameter-free voting law (Fig. 11). Across a $3 \times 4 \times 5$ grid of (p, w, k) (n=3000 each), per-problem majority accuracy against the single scalar $z = \sqrt{k}\, m/\sigma$ collapses onto $\Phi(z)$ with RMSE 0.022 and no free parameters, the residual shrinking as $O(k^{-1/2})$ (mean $|\cdot|$ from 0.028 at k=8 to 0.002 at k=128) rather than sitting at a systematic offset. The margin, not (p, ρ, k) separately, is the sufficient statistic for voting.

P.3 Step 1, the decomposition on real models: protocol and figures

The body's headline table is the frontier and its limits (Table 1); the cross-family replication on the reused pool is Table 8 here. The executed protocol, its relation to the pre-registration, and the two supporting figures follow.

Executed protocol. Data: the *Large Language Monkeys* release (Brown et al., 2024), real repeated samples from Llama-3-8B-Instruct and Llama-3-70B-Instruct on GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) (subset sizes n=127/128), plus Gemma-7B/2B on MATH as a second model family; we subsample k=64 per problem from the provided pool (fixed seed). Correctness oracle: the dataset's validated `is_corrects` labels; the Step-1 answer extractor agrees with them

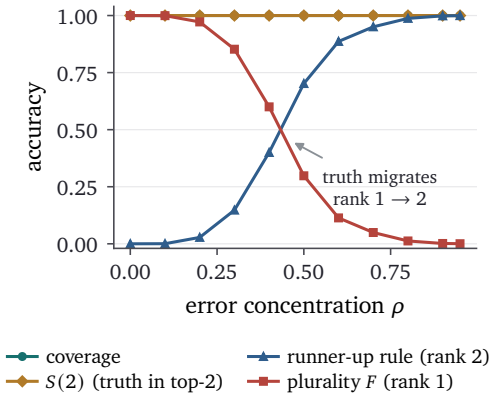


Figure 9: **Voting fails but the truth is only demoted to rank 2.** Same single-dominant sweep. As ρ rises the truth's count-rank moves from 1 to 2 (at $\rho = 0.9$, $\mathbb{P}[\text{rank } 2] = 0.999$): plurality (rank 1) collapses to 0.0015, but the label-free runner-up rule (rank 2) rises to 0.9985, and $S(2) =$ "truth in top-2 counts" $= 1.00$ throughout. Self-consistency is *not* the count ceiling (Prop. F.1); the gap here is count-recoverable.

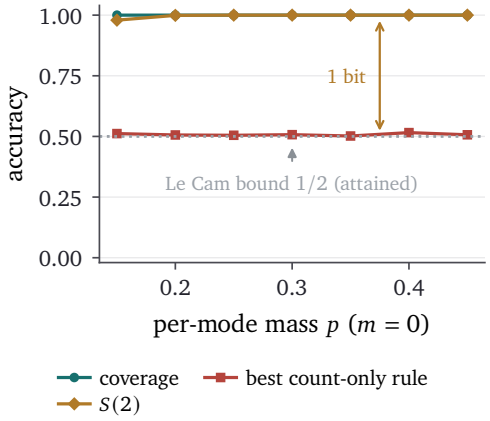


Figure 10: **The attained ceiling: two equal modes ($m = 0$).** When the dominant wrong answer has emission probability equal to the truth's, *no* count-only rule beats the Le Cam bound $1/2$ (best rule 0.51 ± 0.01 across p), yet the truth is in the top-2, so $S(2) \geq 0.98$ (exactly 1.00 for $p \geq 0.25$; 0.979 at p=0.15, where the truth's count is low enough that a *diffuse* wrong answer occasionally takes the second slot). This is the no-free-lunch family of Theorem F.4 and the one-bit closure of Proposition F.8; the single-dominant family (left) does *not* exhibit this collapse.

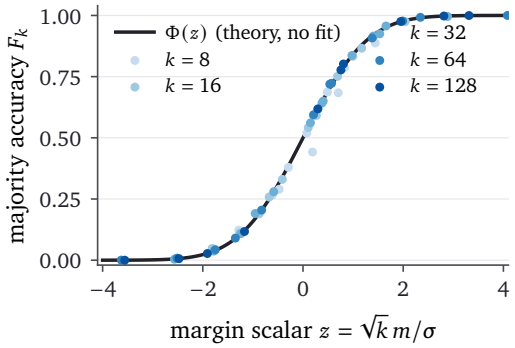


Figure 11: **Parameter-free voting law (Theorem B.3).** Per-problem majority accuracy over a $3 \times 4 \times 5$ grid of (p, w, k) plotted against the single scalar $z = \sqrt{k}\, m/\sigma$, collapses onto the standard-normal CDF $\Phi(z)$ (black) with RMSE 0.022 and no fitted parameters. Points shaded by k; the residual shrinks as $k^{-1/2}$ (Berry–Esseen), not a systematic offset. The margin m, not (p, ρ, k) separately, is the sufficient statistic.

98.9–99.6% on GSM8K and 85–91% on MATH (LaTeX parsing is harder, so the MATH error-*identity* statistics of this step are conservative; §P.5 bounds the effect). That extraction is this step's, and is not the count banks the body's rank measurements are built on: those are rebuilt label-free by the canonicaliser of Appendix D, whose re-grade agrees with the release labels on 98.9–99.96% of samples in every pool cell (Table 2). The two pipelines are reported in separate tables and are never mixed within one. For each cell we compute pass@1, majority-vote accuracy F, the count-only ceiling (best calibrated rank rule on a held-out split), coverage, the list envelope $S(1), S(2), S(4)$, and the gap $\text{Cov}_k - F$, stratified into five difficulty bins by $1 -$ pass@1, with bootstrap 95% CIs over problems (≥ 2000 resamples).

Table 8: **Cross-family replication on the reused pool** (k=64, n=127/128; gap CIs are bootstrap intervals over problems), two further model families and a 70B size. In every cell the gap tracks the rate at which a dominant wrong answer out-votes the truth ("$W{>}a^*$"), the correlated-Condorcet signature. The rank columns are the truth's *observed count* rank at this budget, which Theorem B.2 sends exponentially to its mass rank R^*: mass at rank 2 means the gap is bridgeable by one bit, mass at rank 3+ that it is not. The cell-level GSM8K/MATH difference in the rank-3+ column is the same composition effect stratified away in Appendix E: these MATH cells hold a hard tail these GSM8K cells do not.

Bench	Model	pass@1	F	Cov_k	gap [95% CI]	$W{>}a^*$	rank2	rank3+	$S(2)$
GSM8K	L3-8B	0.77	0.87	1.00	0.13 [0.07, 0.19]	0.13	0.09	0.03	0.97
GSM8K	L3-70B	0.93	0.98	0.99	0.02 [0.00, 0.04]	0.02	0.02	0.00	0.99
MATH	L3-8B	0.27	0.45	0.80	0.35 [0.27, 0.44]	0.43	0.09	0.26	0.54
MATH	L3-70B	0.48	0.60	0.88	0.27 [0.20, 0.35]	0.31	0.12	0.16	0.72
MATH	Gemma-7B	0.21	0.43	0.80	0.38 [0.29, 0.46]	0.45	0.09	0.28	0.52
MATH	Gemma-2B	0.10	0.25	0.59	0.34 [0.26, 0.42]	0.57	0.05	0.29	0.30

Table 9: **The count-recoverable gap at full benchmark scale**; the pool version is Table 11. Columns as there: $\hat{T}_{\mathrm{ct}}$ is the best cross-fitted rank rule, "in-samp." the richest family fitted and scored on the same problems, "headroom" $\mathbb{P}[a^*$ at rank 2 or 3 $\mid$ covered$]$ with the count of profile cells preferring a rank other than 1.

Bench	Model	F	rank-2 everywh.	restricted rank acc. [CI on Δ]	in-samp.	headroom (cells $\neq$ 1)
GSM8K	Qwen2.5-3B	0.907	0.390	0.907 [0.000, +0.000]	0.910	0.055 (2/19)
MATH-500	Qwen2.5-3B	0.724	0.212	0.724 [0.000, +0.000]	0.728	0.122 (2/18)
GSM8K	Qwen2.5-7B	0.925	0.672	0.925 [0.000, +0.000]	0.929	0.049 (5/19)
MATH-500	Qwen2.5-7B	0.768	0.362	0.774 [−0.002, +0.016]	0.780	0.105 (4/21)

Table 10: **Step 3 at full benchmark scale.** Per cell: the diagnostic-slice size, the count-only rank-2 baseline on the same problems, the best-scoring member of the pre-registered seven-member family, its shortfall against that baseline, and the improvement the test could have detected at 80% power under the Holm-corrected α. No cell produced a rejection.

Bench	Model	slice n	count base	best member	acc	shortfall	detectable $+\delta$
GSM8K	Qwen2.5-3B	102	0.578	self-evaluation	0.216	−0.363	+0.160
MATH-500	Qwen2.5-3B	96	0.354	self-evaluation	0.269	−0.085	+0.175
GSM8K	Qwen2.5-7B	76	0.737	self-evaluation	0.316	−0.421	+0.155
MATH-500	Qwen2.5-7B	84	0.452	self-evaluation	0.185	−0.267	+0.185

Relation to the pre-registration. The plan specified generating our own k=64, T=0.8 samples on full GSM8K (1319) and MATH-500 with `math_verify` checking. Step 1 ran before a sampling backend was available, so it was executed on the reused pool instead (same benchmarks, same family for one model, smaller n), with the deviation logged before outcomes were seen; the grid cells of §3 later executed the original specification, and both are reported. The *predictions* were fixed unchanged throughout. **P1:** the gap is small on easy problems and large on hard ones, monotone in difficulty. **P1′:** report the truth's count-rank distribution on the diagnostic slice, whether the gap is bridgeable by one bit (mass at rank 2) or not (mass spread deeper), and whether $S(2) \approx \mathrm{Cov}_k$ as the top-two theorem predicts.

P.4 Step 1c, the count-only ceiling at dataset level

Why this measurement is needed. It estimates a quantity the theory already names rather than testing a new prediction, and it was added after Step 1 outcomes were known (logged in the pre-registration as post-hoc). Table 8 shows the truth at count-rank 2 on a large share of the failure slice, and §P.6 shows a rank-2 rule recovering 80% of that slice. Neither is T_{ct}. Both condition on the diagnostic slice, which is defined by the gold label and cannot be identified without one; a rank rule that must run on *every* problem buys the slice's accuracy only if the sorted count profile tells it when to depart

Table 11: **The headroom above the plurality is real, and no count-profile rule we fit locates it** (pool cells plus the own-sampled pilot cell). $\hat{T}_{\mathrm{ct}}$ is the best cross-fitted rank rule over five families, an attainable *lower* bound on the count ceiling; "in-sample" is the richest family fitted and scored on the same problems (an optimistic anchor, not an estimate). "rank-2 everywhere" is the naive reading of the runner-up phenomenon deployed without a regime detector. "headroom" is $\mathbb{P}[a^*$ at rank 2 or 3 $\mid$ covered$]$, the mass a rank rule competes for; "cells $\neq 1$" is how many of the fitted profile cells prefer a rank other than 1, and the share of problems they hold. The full-scale version is Table 9.

Bench	Model	F	rank-2 everywh.	$\hat{T}_{\mathrm{ct}}$ [CI on Δ]	in-samp.	headroom	cells $\neq 1$ (% probs)
GSM8K	L3-8B	0.874	0.339	0.874 [−0.00, +0.00]	0.882	0.102	1/16 (0.8)
GSM8K	L3-70B	0.976	0.622	0.976 [−0.00, +0.00]	0.976	0.016	0/7 (0.0)
MATH	L3-8B	0.445	0.102	0.445 [−0.00, +0.00]	0.461	0.196	2/17 (3.1)
MATH	L3-70B	0.602	0.203	0.602 [−0.00, +0.00]	0.617	0.152	1/14 (3.1)
MATH	Gemma-7B	0.430	0.094	0.453 [−0.03, +0.08]	0.469	0.194	3/12 (28.9)
MATH	Gemma-2B	0.250	0.047	0.250 [−0.00, +0.00]	0.250	0.120	0/10 (0.0)
GSM8K	ours-3B	0.867	0.187	0.867 [−0.00, +0.00]	0.880	0.048	2/17 (1.3)

from rank 1. Whether it does is the empirical question, and the decomposition (2) needs its answer, not the slice's.

Protocol. By Proposition F.1 the Bayes count-only selector is a rank rule, a map from the sorted count profile to a count-rank. We fit the plug-in version: discretize the profile by statistics the theory names as regime indicators (the top-1 share $N_{(1)}/k$, the runner-up ratio $N_{(2)}/N_{(1)}$, and the number of distinct answers, with bin edges fixed a priori), and in each cell output the rank at which the truth most often sits. Five families of increasing capacity are fitted, from a single constant rank up to a 48-cell rule. Calibration reads gold labels, which is legitimate here and only here: T_{ct} is a *ceiling*, the value of a Bayes rule that Theorem B.1(iii)'s scope note already calls non-constructive in the prior, not a deployable selector. To keep the estimate from being an in-sample fantasy we **cross-fit 5-fold over problems**, so no problem contributes to the rule that scores it, and we report the in-sample fit beside it as an optimistic upper anchor. Differences from plurality carry a paired bootstrap CI over problems.

Result: the count-recoverable term measures to zero, at pilot and at full scale. Table 11. On six of seven cells the cross-fitted rank rule reproduces plurality *exactly*, so $\hat{T}_{\mathrm{ct}} - F = 0$; on the seventh (Gemma-7B) it gains 0.023, with a paired bootstrap CI $[-0.031, +0.078]$ containing zero. The result is not an artifact of the fold split or of estimator variance, because it survives the most generous possible version of the test: fitted *in-sample* on the richest 48-cell family, with every advantage given and no honesty imposed, the gain over plurality is at most 0.039 and typically ≤ 0.016. Deploying rank 2 everywhere, the naive reading of Remark F.2, is catastrophic, costing 0.20–0.68 accuracy, which is the arithmetic the slice-conditional numbers conceal.

The same measurement at full benchmark scale (Table 9, n=1319 and n=500 rather than $n \approx 128$) reproduces this exactly. On three of the four grid cells the cross-fitted rank rule equals plurality to three decimal places; on MATH-500/7B it gains 0.006 with a paired interval $[-0.002, +0.016]$ containing zero, and the in-sample anchor adds at most 0.006 on top of that. The headroom is again real and again unlocated: the truth sits at count-rank 2 or 3 on 4.9–12.2% of covered problems, and of the 18–21 populated profile cells per grid cell, between two and five prefer any rank other than 1. The larger n removes the obvious reading of the pool result: that coarse rules at $n \approx 128$ were simply underpowered.

Why, precisely. The last two columns separate the two candidate explanations, and it is the second. The headroom is real: the truth sits at count-rank 2 or 3 on 1.6–19.6% of covered problems, so Theorem B.1's chain is not collapsed and the information is not lost to the counts. But of the 7–17 populated profile cells per dataset, between 0 and 3 prefer any rank other than 1, and outside the Gemma-7B cell those hold under 3.1% of problems. The sorted count profile simply does not say which regime a problem is in. This is exactly what the local ceiling law predicts: on the labelling-symmetric two-mode family the Bayes count rule is a sign test whose *sign is supplied by the prior*,

not by the sample; a dataset that mixes $m > 0$ and $m < 0$ problems with similar count profiles gives a count-only rule no way to recover the sign.

What this changes. Proposition F.1 is untouched: it is a theorem, and the synthetic single-dominant family attains it (Fig. 9, rank-2 scoring 0.9985 where plurality scores 0.0015). What the measurement withdraws is the empirical reading that a better count rule is a cheap route across the gap on real model–task pairs. It is not, on any cell we have. Two consequences follow. First, the honest baseline for a trace-aware selector is numerically F after all, because $\hat{T}_{\mathrm{ct}}$ coincides with it; the selector sections below are unaffected, since they compare against the slice-conditional rank-2 rule, which is the *stricter* of the two. Second, external verification bits are the only route across the gap this paper demonstrates rather than merely permits, which sharpens §9 from a menu into a recommendation.

How far this reaches. These rule families discretize three theory-named regime indicators, and the richer gradient-boosted family of §6 returns the same verdict on all ten cells: *these* count statistics do not separate the $m > 0$ and $m < 0$ regimes at these sample sizes. Proposition 2.3 names what a statistic that did would have to be, a detector of rank miscalibration, which is the quantity we estimate directly rather than hunt for by bake-off.

P.5 Step 2, does error correlation explain the gap?

Pre-registered protocol. Per problem we estimate the error concentration ρ (unbiased collision estimator over wrong samples), the voting-failure rate, and the rate at which the dominant wrong answer out-votes the truth. We fit the two-mode prediction $\hat{F}(p, \rho, k)$ from §2 and compare against a p-only baseline. **Prediction P2:** the correlation term materially improves the fit (report R^2 and likelihood ratio with CIs), and voting failure rises with ρ. **P2′ (pre-registered addition):** test the parameter-free scaling collapse of Theorem B.3, per-problem majority accuracy against the single scalar $\sqrt{k}\,\hat{m}/\hat{\sigma}$ should lie on the standard-normal CDF with no fitted parameters; measure $\hat{w}$ directly (the margin, not ρ, is the primary mechanism variable). **Stop condition (pre-registered):** if neither ρ nor the margin explains the gap beyond p, we reject the proposed mechanism; the Step 1 measurement result remains independently supported.

A circularity, and how we avoid it. A naive test of Theorem B.3 estimates the margin $\hat{m}$ from the *same* k samples that decide the vote; then $\mathrm{sign}(z)$ equals the vote outcome almost by construction (in the two-mode reduction, majority-correct $\Leftrightarrow N_{a^*} > N_W \Leftrightarrow \hat{m} > 0$), so an apparent "perfect classification" is an identity, not a confirmation. Our in-sample statistics show exactly this artifact, a step function at $z{=}0$ (Fig. 12, open markers; decisive-region sign-accuracy 1.00 by construction). We therefore test the law *out of sample*: estimate $(\hat{m}, \hat{\sigma})$ on one random half of each problem's samples and predict the majority vote on the *disjoint* other half.

Result (real data). Held out, the margin law survives de-circularization (Fig. 12, filled markers; $n{=}1075$ problem-splits pooled over all six cells). The predicted $\Phi(z)$ tracks the held-out majority-vote accuracy with calibration error $\overline{|\mathrm{emp} - \Phi|} = 0.045$ (GSM8K 0.066, MATH 0.050), and the *sign* of the training-half margin predicts the held-out vote with accuracy 0.93 on the decisive slice ($|z| \geq 0.5$), versus the tautological 1.00 in-sample. Thus the margin, not ρ or p alone, is a genuine, out-of-sample predictor of voting on real data.

The transition-region dip is partly predicted, not merely excused. The held-out curve sits below Φ in the transition region $z \in [0.5, 2)$. Attributing this to estimation noise in $\hat{m}$ at half sample size names a mechanism without predicting anything. The lattice expansion does predict something: because the contest walk has span-1 integer steps, the one-term Edgeworth continuity correction contributes $-\phi(z)/(2\sigma\sqrt{k})$, which is *negative at every* z. We test that sign directly: computing the exact lattice contest probability by convolution at each problem's plug-in $(\hat{p}, \hat{q})$ and comparing it to $\Phi(z)$, the discreteness deviation is negative in 8 of 8 z-bins, as predicted, and the closed form tracks the exact lattice value to ≤ 0.011 in every bin. Its magnitude, however, accounts for only part of the effect: in $z \in [0.5, 2)$ the lattice correction is -0.037 against an observed -0.236, i.e. 16%. We therefore report the decomposition rather than either explanation alone: the dip's *sign and location* are a predicted lattice effect, and the residual -0.199 is estimation noise in $\hat{m}$, which is the errors-in-variables flattening one expects when the binning variable is itself estimated at half sample size.

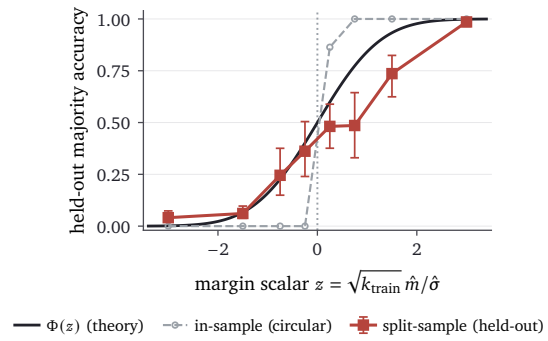


Figure 12: **De-circularized margin law (P2′).** Held-out majority-vote accuracy against the margin scalar $z = \sqrt{k_{\text{train}}}\,\hat{m}/\hat{\sigma}$ (pooled). The *in-sample* statistic (grey, open), which takes margin and vote from the same samples, is a step function at $z{=}0$: this is the circularity, not a confirmation. The *split-sample* statistic (red) estimates $(\hat{m}, \hat{\sigma})$ on one half and predicts the vote on the disjoint other half; it tracks the theory curve $\Phi(z)$ (black) out of sample (calibration error ≈ 0.05; decisive-slice sign-accuracy 0.93). This is a genuine test of Theorem B.3.

Concentration monotonicity, on real data. That more concentrated errors lower F at fixed p (a theorem on the single-dominant family) was validated only synthetically. We test it on the real pool with a rank statistic robust to ρ's scale: among covered problems matched within a 0.1-wide pass@1 band, we ask how often the *more concentrated* of a voting-discordant pair is the one that fails. Across all cells with resolvable pairs this concordance is 0.89–0.98 (chance 0.5; GSM8K-8B 0.98 over 47 pairs, MATH-8B 0.93 over 308, Gemma-2B 0.89 over 398). At matched accuracy, higher error concentration predicts voting failure on real data, which confirms the mechanism beyond the margin itself.

What P2 was and was not tested by. The pre-registered comparison of a full per-problem $\hat{F}(p, \rho, k)$ fit against a p-only baseline was superseded rather than run: the split-sample margin law (P2′) tests the same mechanism out of sample and without the circularity that a same-sample likelihood ratio would inherit, and the matched-difficulty concordance test above supplies the ρ direction. We report P2 as answered by those two rather than by the originally specified regression, and record the substitution here rather than in a footnote.

Extractor-noise sensitivity (falsification test). MATH error-*identity* statistics rest on an extractor that disagrees with the validated correctness labels on $\epsilon = 9$–15% of samples per cell. If the correlated-error mechanism were an artifact of that noise, perturbing wrong-answer identities at rate ϵ should move the headline statistics by comparable amounts. It does not (8 seeds): *fragmenting* identities at rate ϵ (the direction of real parse misses) shifts the $W{>}a^*$ rate by only -0.02 to -0.04 against reported effects of 0.31–0.57, quantifying, rather than asserting, that the published numbers are conservative, while even *adversarially merging* identities at rate ϵ shifts it by $+0.03$ to $+0.15$, leaving every qualitative claim intact; $S(2)$ moves by ≤ 0.03 under both perturbations. The mechanism is not an extractor artifact.

P.6 Step 3, trace-aware selectors on the diagnostic slice

Pre-registered protocol. The diagnostic slice is the set of problems that are covered but majority-vote-wrong, by construction the only place a better label-free selector can improve on voting. Three candidate feature maps Φ, fixed in advance: (a) structural diversity of independent derivations per answer (cluster traces by intermediate quantities; prefer the answer supported by the most distinct valid derivations); (b) aggregated label-free self-evaluation; (c) an intermediate-agreement / cross-consistency graph score; and (d, P3′, exploratory) a surprisingly-popular / second-order feature, select the answer whose realized frequency most exceeds the model's own predicted frequency, whose sharp null (no gain exactly when the meta-prediction matches the first-order distribution) measures second-order self-knowledge on hard problems. **Primary metric:** because the slice is a binary top-two discrimination (Theorem F.7), chance is exactly $1/2$; each Φ must beat *both* $1/2$ and the count-only ceiling $T(\Phi_{\text{counts}})$, with bootstrap CIs and Holm–Bonferroni correction across the feature

list. **Prediction P3:** at least one Φ beats the count-only ceiling materially. **Both outcomes are results:** $T_{\mathrm{tr}} \gg T_{\mathrm{ct}}$ means trace signal exists and selection research has a target; $T_{\mathrm{tr}} \approx T_{\mathrm{ct}}$ bounds what verification-free selection reaches on these features. We run the family at full benchmark scale, with a power analysis, and then measure the feature-level quantity that bounds every rule the features generate.

Result: the full family, at scale, with stated power. The grid cells carry their own log-probabilities and second-order elicitations, so all seven pre-registered members run, including surprisingly-popular (Prelec et al., 2017) and confidence-weighted voting (Tian et al., 2023). Table 10 reports each cell's diagnostic slice (n = 76–102), the count-only rank-2 baseline on those same problems, and the best-scoring of the seven. The best member is self-evaluation on all four cells, and it falls 0.085 to 0.421 *below* the count baseline; the tests carried 80% power under the Holm-corrected α against improvements of $+0.155$ to $+0.185$, so a gain of the size P3 anticipated would have been detected on every cell and appeared on none. The negative is therefore bounded rather than merely unrejected: none of these seven features helps by more than about 0.16 here.

The pool slice, and what it settles. The same protocol on the pool runs only the two *model-free* features, the release carrying neither log-probabilities nor second-order elicitations. On the GSM8K diagnostic slice (covered but voting-wrong, both models pooled, n=20, k=128), against chance $1/2$: plurality 0.00 by construction, **rank-2 (count-only ceiling)** 0.80 **[**$0.60, 0.95$**]**, structural diversity 0.30 $[0.10, 0.50]$, diversity-per-sample 0.20, intermediate agreement 0.15 $[0.00, 0.30]$. What this slice establishes is the count-rank result: the rank-2 rule recovers 80% of it, confirming the truth sits at count-rank 2 (§2, Table 8). The trace comparison on 20 decisions is not powered against the alternatives that matter, and no conclusion here rests on it; the grid run above is the instrument for that question.

The feature-level bound. §P.7 measures the quantity Theorem F.5 proves is decisive, the per-sample mode affinity, estimated from every sample rather than from a handful of selector decisions, which bounds the entire selector family a feature generates rather than the one rule we happened to code.

P.7 Step 3c, measuring the quantity the theorem says is decisive

A selector bake-off on a small slice cannot settle whether trace features carry signal. Theorem F.5 lets us stop running bake-offs and measure the quantity that decides the question instead: the per-sample mode affinity ρ. This is a better instrument for the same reason a power calculation beats a failed experiment: ρ is a property of the *feature*, estimated from all 8192 samples of the slice rather than from 14 selector decisions, and it bounds the entire selector family the feature generates rather than the one rule we happened to code. This subsection develops the instrument on an own-sampled GSM8K pilot cell, whose diagnostic slice holds 14 problems; the full-scale version, run on all four grid cells, is the right-hand panel of Figure 14.

Protocol. On the own-sampled diagnostic slice, for each per-sample statistic we form its values in the truth cluster and in the dominant wrong cluster, center per problem (the affinity is invariant to a shift applied to both laws), and estimate ρ by a Gaussian fit, with a histogram plug-in as a binning-free cross-check. The null is exactly the mode swap τ of Theorem F.4: relabel which cluster is the truth, independently per problem, 4000 times.

A confound that manufactures significance, and must be removed. On the diagnostic slice the truth cluster is smaller than the dominant wrong cluster on 14 of 14 problems, not as an empirical finding but *by the definition* of "majority vote was wrong." Any statistic correlated with cluster size therefore shows apparent mode separation that is pure bookkeeping. Both structural statistics we test are so correlated (-0.26 with log cluster size), and the naive estimate duly reports strong separation, $\rho = 0.877$ ($p < 0.001$) for fingerprint size and $\rho = 0.960$ ($p = 0.003$) for trace length, which would read as a positive result. It is not one: "the truth's traces are longer" here reduces to "smaller clusters have longer traces," which is collinear with the count and so buys a selector nothing beyond the count-rank it already had. We therefore residualize every statistic on log cluster size before estimating ρ. Where size and role are perfectly collinear the two explanations are not separately identifiable and residualization absorbs genuine signal too, so the reported ρ is *conservative*: it can understate separation, never overstate it.

Table 12: **Per-sample mode affinity ρ of each trace feature** (own-sampled GSM8K cell, diagnostic slice n=14, ρ estimated on the 13 problems with a non-degenerate live cluster). $\rho = 1$ means the feature cannot break the mode symmetry at any k (Theorem F.5(i)). All three sit inside the mode-swap null's 95% band. "LLR" is the decoder the theorem says is optimal, Gaussian-calibrated leave-one-problem-out; none beats the count-only rank-2 rule (0.286 on these problems), which is what a null ρ predicts.

Feature	ρ	null 95% band	p	T_{tr} at k'=34	LLR acc.
mean token log-prob	0.9985	[0.989, 1.000]	0.675	[0.52, 0.65]	0.214
trace length	0.9931	[0.980, 1.000]	0.328	[0.61, 0.81]	0.214
fingerprint size	0.9625	[0.942, 0.998]	0.195	[0.86, 0.98]	0.214

Result. Table 12: no feature departs from the no-separation null ($p = 0.195$–0.675). The direct check agrees. A leave-one-problem-out calibrated log-likelihood-ratio decoder, the rule Theorem F.5 says would cash any real asymmetry, scores 0.214 on the slice for all three features, *below* the count-only rank-2 baseline of 0.286. The realized accuracy also falls far below the affinity bounds, which is itself informative: it is the homogeneity failure flagged after Theorem F.5, since (P_c, P_w) estimated by pooling are misspecified for individual problems.

What this licenses. This is a strictly stronger negative than a selector null, on two counts: it is a statement about the *features*, which bounds every rule they generate rather than the three we coded, and a measured ρ does not drift toward significance as the slice grows. Its reach is one model, one benchmark, three features, and 13 exchangeable units in the permutation null (one of the 14 slice problems has a live cluster too small to estimate an affinity and drops out), which resolves gross separation and nothing subtle. Read precisely: these features carry no usable mode asymmetry here, and none is claimed to exist or not exist elsewhere. Theorem F.5(ii) is the reason to keep looking rather than to stop, since a per-sample edge far too small to see in any bake-off is still amplified exponentially in k, which is what makes the feature-level instrument the right one to scale up, which Figure 14 does: at full benchmark scale, across four cells, every feature but two sits inside its own mode-swap null band.

This pilot cell supports no controlled comparison against a pool cell: its error concentration, $\rho = 0.2075$, matches no pool cell, and none is drawn from it. It is used here for the instrument and nowhere else.

P.8 The count-ranked verification protocols in full

Table 13 is the full version of the one-bit figure quoted in §3: every S_{2^B} on every grid cell, so a reader can see what the *later* bits are worth as well as the first.

Table 13: **What a count-ranked shortlist plus B oracle bits achieves** (grid cells, k=64; bracket is a bootstrap 95% CI over problems). The columns report S_{2^B}, the *count-ranked* construction of Corollary 2.5*(ii)*: shortlist the top 2^B answers by count and spend B noiseless membership bits inside it. These are never larger than the unrestricted Bayes values $G_{2^B}(\Phi_{\mathrm{counts}})$ of Theorem B.1(iii), which they lower-bound, and equal them under Corollary 2.4. The first bit buys more than the second on every cell, but what separates these cells is what the later bits are still worth: on GSM8K bits 2–4 add only 0.016–0.029, the curve being all but exhausted, and on GSM8K/7B four bits have already reached coverage exactly; on MATH-500 the same three bits add 0.086–0.112, more than the first bit did. Diminishing returns per *bit* are therefore not automatic: Theorem B.1(v) gives concavity per list *slot*, and a bit doubles the slot count. As everywhere in this paper, the cell-level contrast is a contrast of difficulty mixes (Appendix E), not of benchmark identity. The final column is computed by the grid decomposition pipeline and its headline counterpart in Table 1 by the rank pipeline; the two agree within every bootstrap interval, differing by at most 0.04.

Bench	Model	F (B=0)	$S(2)$ (B=1)	$S(4)$ (B=2)	$S(16)$ (B=4)	Cov_k	gap closed by 1 bit
GSM8K	Qwen2.5-3B	0.907	0.952	0.967	0.981	0.985	0.578 [0.48, 0.68]
MATH-500	Qwen2.5-3B	0.724	0.792	0.858	0.904	0.916	0.354 [0.26, 0.45]
GSM8K	Qwen2.5-7B	0.925	0.967	0.975	0.983	0.983	0.737 [0.64, 0.84]
MATH-500	Qwen2.5-7B	0.768	0.844	0.884	0.930	0.936	0.452 [0.34, 0.56]

P.9 GRID FIGURE PANELS

The remaining full-scale panels.

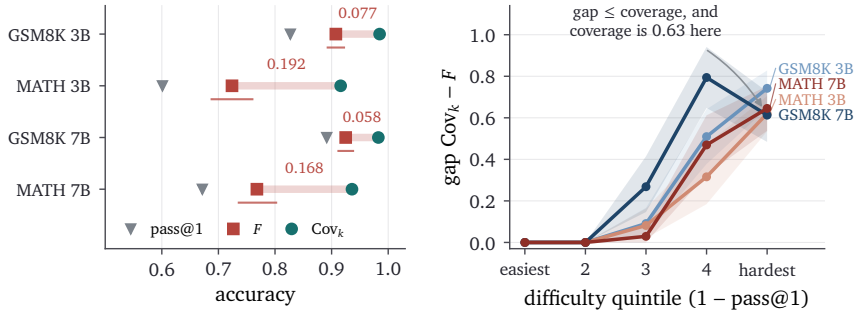


Figure 13: **Where each cell sits, and why (P1).** *Left:* pass@1, self-consistency F, and coverage per cell, with the gap bracketed and its bootstrap CI ruled beneath. The model-size step moves F toward a coverage that barely moves: the gap closes from below, by selecting better, not by covering less. *Right:* the gap against difficulty (five equal-width bins of $1 - \text{pass@1}$; bands are per-bin bootstrap CIs). The gap is exactly zero on the two easiest quintiles of every cell and rises steeply thereafter. The raw gap is strictly monotone on three of four cells; on GSM8K/7B it turns down in the hardest quintile because coverage itself has fallen to 0.629 there, and dividing by coverage restores strict monotonicity on all four cells (hardest quintile 0.92–0.98). Difficulty is estimated from the same k samples, which biases per-bin *levels*. The trend, which is the P1 claim, does not depend on that: cross-fitting the difficulty estimate on the count banks (Appendix E.1) leaves the coverage-normalised gap non-increasing in the bin's held-out pass@1 on 10 of 10 cells.

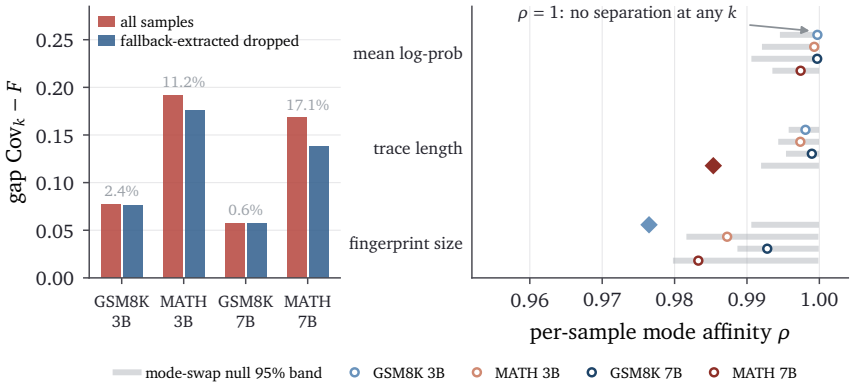


Figure 14: **Two threats to the grid, measured rather than argued.** *Left:* the gap with all samples against the gap after dropping every sample whose answer came from the fallback extractor (percentages above each pair are that cell's fallback rate). The gap survives everywhere and the benchmark ordering is unchanged; the one material movement is MATH-500/7B ($0.168 \to 0.138$), cited alongside the headline wherever that cell's magnitude is load-bearing. *Right:* the per-sample mode affinity ρ of each trace feature against its own mode-swap permutation null (95% band in grey). By Theorem F.5 a feature with $\rho = 1$ cannot break the truth/wrong-mode symmetry at *any* k, so this panel bounds whole selector families rather than individual rules. Filled diamonds mark the two feature–cell pairs that fall below their null band; every other point sits inside it.

P.10 THINKING-MODE INTERVENTION

The preceding cells predict the frontier from a fixed model's rank distribution. We additionally intervene on that distribution. A frozen, paired design runs the pinned Qwen3-8B-FP8 checkpoint (Qwen Team, 2025) on 250 preregistered MATH-500 problems with 16 samples per arm. The

checkpoint, prompt bytes, temperature, top-p, top-k, token cap, extractor and grader are identical; only the chat template's `enable_thinking` boolean changes. The arms use disjoint seed streams. All 8,000 generations completed, and all intervals below are paired problem-level bootstrap intervals.

Thinking raises pass@1 from 0.8558 to 0.9193, a difference of $+0.0635\,[+0.0333, +0.0935]$, while coverage at $k{=}16$ changes by $-0.016\,[-0.036, 0.000]$. The intervention instead moves the truth toward the top of the answer ordering: $\mathbb{P}[R^* \leq 1]$ rises from 0.9042 to 0.9560, $+0.0518\,[+0.0160, +0.0896]$, and the coverage–selection gap falls by $0.0676\,[0.0320, 0.1038]$. This is preregistered outcome class C: the rank distribution changes without a detectable coverage increase.

The key test uses disjoint data. On repeated random halves, half A estimates the answer-mass ordering and predicts the arm contrast; half B measures it on fresh samples at $k{=}8$. Because an eight-sample plug-in distribution is over-concentrated, the discrepancy is calibrated against a parametric null that resamples from each problem's fitted answer mass, where the rank law holds exactly. Figure 15 shows that the observed discrepancies are ordinary under that null: $p = 0.490, 0.460, 0.725$ for $L = 1, 2, 4$. Thus the intervention's movement of the independently measured selection frontier is quantitatively accounted for by its movement of the truth-rank distribution.

The contrast is internally controlled but not universal. It uses the official FP8 rather than bf16 checkpoint; non-thinking retains the thinking arm's decoding parameters to keep a single-variable intervention; truncations are abstentions and occur more often with thinking (6.5% versus 1.2%); and the split-bank halves contain only eight samples. None changes the paired claim, but each limits extrapolation beyond this checkpoint and task.

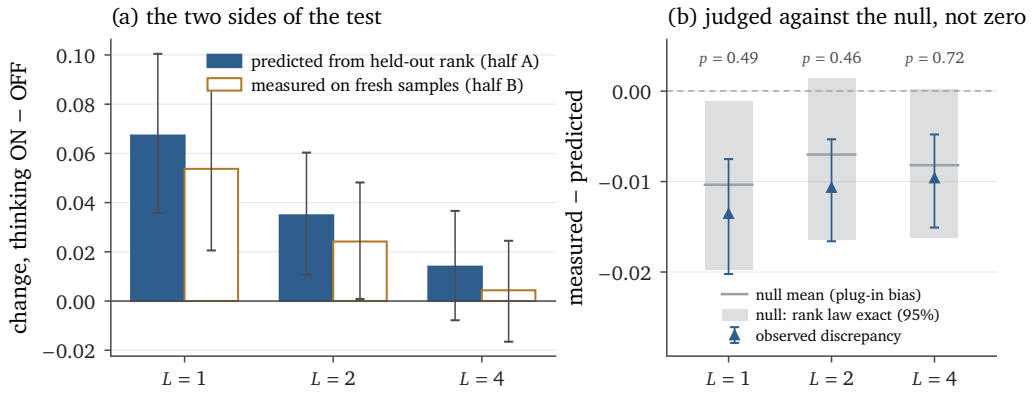


Figure 15: **The rank law under intervention. (a)** The change from thinking OFF to ON predicted from one half-bank against the change measured on the other, at $k{=}8$; bars are paired 95% bootstrap intervals. **(b)** The quantity the verdict is read from: the discrepancy between those two sides, against the parametric null (grey band, 95%) in which the rank law holds exactly. The null is displaced from zero because an eight-sample plug-in distribution is over-concentrated, so the estimator returns a negative discrepancy even when the law is exact; the observed discrepancy falls inside that band at every L. Judging panel (a) by whether its intervals overlap instead reproduces the uncalibrated test, which rejects at every L and is not the verdict.