When More Sampling Hurts: The Modal Ceiling and Correlation Ceiling of Test-Time Scaling

Yong Yi Bay* **Kathleen A. Yearick***

PhD, University of Illinois at Urbana-Champaign

ABSTRACT

People overthink; language models over-sample, and the extra effort can talk both into a *worse* answer. Reasoning systems answer a hard question by sampling it many times (*test-time scaling*), and the more they draw, the more often a correct answer turns up somewhere, so *coverage*, the fraction of problems with at least one correct try, climbs and appears to be progress. But a deployed system must return one answer, and choosing it, not knowing which try is right, is *selection*; selection is capped, and past a point extra samples only make the model surer of a confident mistake, even as every draw adds cost. The gap between climbing coverage and stalled selection, the *identifiability gap*, is the answer a model can produce but not pick. So the real question is not whether to sample but how far, and the answer is: not far. For picking an answer, the vote has already settled within a few dozen draws, the *modal ceiling*; for scoring a benchmark, sooner still, the *correlation ceiling*. Beyond that, extra draws cost compute and add nothing, and can even make the answer worse. This paper turns the cutoff into a single number, the *effective number of samples*, that any sampling run already reveals. The bottleneck is recognizing a right answer, not generating one.

Keywords test-time scaling · inference-time compute · repeated sampling · pass@k · coverage · best-of-n · self-consistency · effective sample size · design effect · intraclass correlation · correlation ceiling · modal ceiling · identifiability gap

1 Introduction and roadmap

A reasoning model can be made stronger at inference time by spending more compute, and the compute goes to one of two levers: longer reasoning, letting the model think more before it answers, or more sampling, drawing n answers to one prompt and combining them [1–4]. This is test-time scaling, and it drives much of the progress in reasoning systems [5–8]; the sampling lever is the subject here. Its headline curve is *coverage*, the fraction of problems on which at least one of n samples is correct [9], and it climbs over several orders of magnitude of n [2], reading as steadily rising capability. That reading is too generous. A deployed system must return a single answer, and with nothing to certify which sample is right it can only *select* one, by frequency or a learned score; selection is capped, and more sampling does not lift it and can even lower it. A correct answer thus grows ever easier to reach but no easier to return: the bottleneck in test-time scaling is not generating a right answer, it is recognizing one. And because each draw costs compute, the question that matters is how far to sample, which the ceilings below answer with a small, goal-dependent budget (Figure 1).

*Equal contribution. Correspondence: {yongyibay, kallie.a.yearick}@gmail.com.

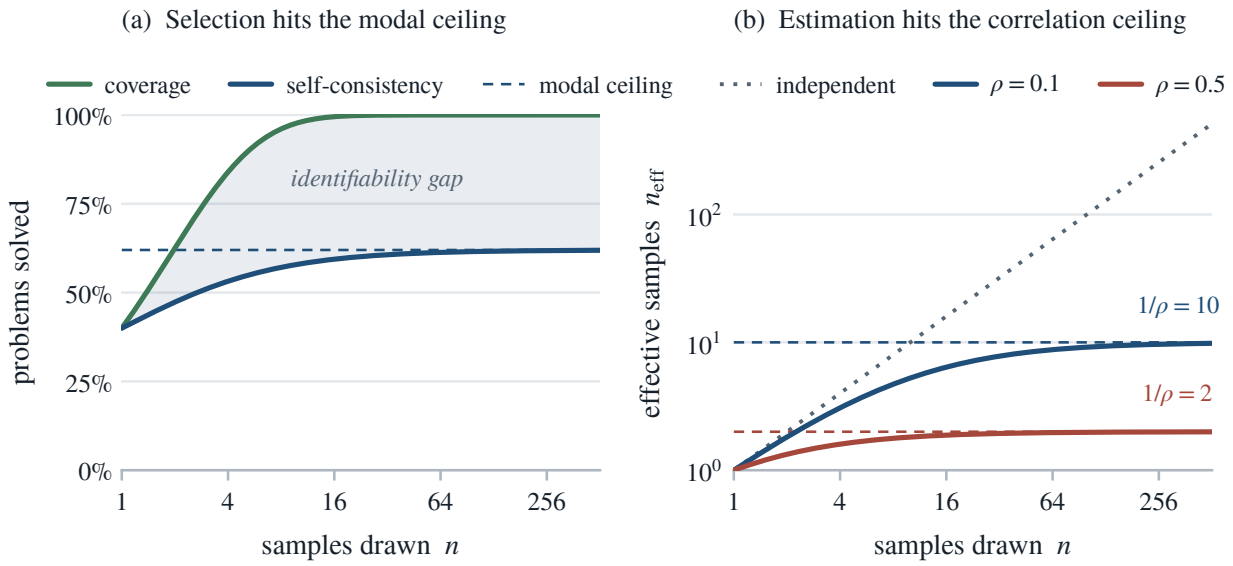

Figure 1: The two ceilings of test-time sampling. **(a)** Coverage, self-consistency, and the identifiability gap between them. **(b)** The effective number of samples n_{eff} against the nominal count n, with the correlation ceiling $1/\rho$.

The split. Coverage and selection ask different questions of the same samples, and only one of them keeps improving. Coverage asks whether *any* sample is correct; a verifier that could spot the correct one would let coverage ride past every limit toward what the model can ever reach. Selection must commit to one answer with no such verifier, and a plurality vote converges to the model's most common answer; on the problems where that answer is wrong, more samples only make the wrong answer win more surely, so selection saturates at the fraction of problems whose most common answer is correct, and can even decline as the budget grows. Brown et al. [2] report exactly this pattern without a mechanism, that "majority voting and reward models plateau beyond several hundred samples and fail to fully scale with the sample budget." The distance between rising coverage and stalled selection is the set of problems the model can solve but cannot pick out, the *identifiability gap*: it can generate the right answer without being able to select it (Figure 2).

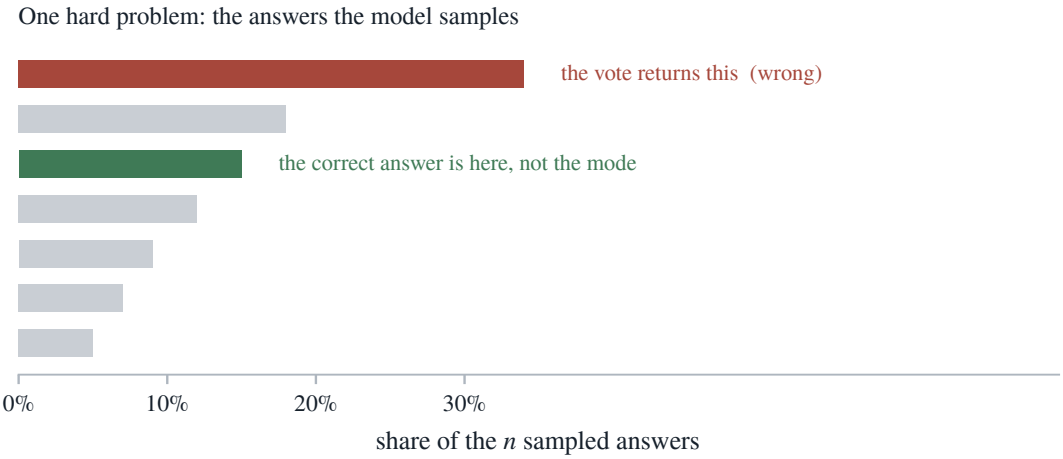

Figure 2: Generated but not selected. On a hard problem the correct answer appears among the sampled attempts, so coverage finds it; but it is not the most common answer, so the plurality vote returns a confident wrong one.

A second, separate ceiling. Even the benchmark accuracy that sampling is meant to estimate is worth less than the sample count suggests, for a different reason. The n attempts to a problem are not n independent tries; they are a cluster drawn from one prompt, alike the way several people from one household are alike in a survey. The classical correction is the *design effect* $d_{\text{eff}} = 1 + (n-1)\rho$ of Kish [10], where ρ is the intraclass correlation among the attempts, so the n correlated draws are worth only n/d_{eff} independent ones [11]. Read for test-time sampling, this is the *effective number of samples*

$$n_{\text{eff}} = \frac{n}{1+(n-1)\rho}, \qquad n_{\text{eff}} \longrightarrow \frac{1}{\rho} \quad \text{as } n \to \infty, \tag{1}$$

which saturates at a hard *correlation ceiling* $1/\rho$: no budget makes n correlated attempts worth more than $1/\rho$ independent ones (Section 2.1). This ceiling governs estimation, not selection; the two are kept apart throughout. The identifiability gap is measured in Section 4.3.

What is and is not new. The design effect and effective sample size are not new here; they are Kish [10] and Cochran [11], and the correction of Condorcet's theorem for correlated voters is older still [12, 13]. Three recent works apply the same instrument to language-model outputs, but to different objects: Kohli [14] measure the effective votes of a panel of *distinct judge models* in evaluation, Goel et al. [15] quantify error correlation *across models*, and Nitarach [16], in a competition report, track the effective sample size of majority voting across mixed models. None treats single-model test-time scaling as cluster sampling, separates the estimation ceiling from a distinct selection ceiling, derives the modal-answer ceiling and its anti-scaling, or connects to the survey-sampling literature. On the coverage side, Schaeffer et al. [17] and Kazdan et al. [18] explain the power-law shape of coverage by a heavy-tailed distribution of *per-problem* difficulty, and Levi [19] reaches a power law through a memorization ansatz; all assume attempts are conditionally independent given the problem, the within-problem $\rho_w = 0$ case of the model below, from which the analysis here recovers a power law of the same form. The contribution here is the lens that separates the three quantities a nominal sample count confounds: a between-problem difficulty spread ρ_b that caps benchmark estimation and shapes coverage, a within-problem dependence measured to be near zero, and the concentration of the answer distribution that caps selection at a modal-hit rate π_{mode} and makes it anti-scale, all grounded on released logs.

This work is a citable derivation, not a new algorithm. Section 2 sets up test-time sampling as cluster sampling and fences the scope of the exact claims. Section 3 derives the effective number of samples, the correlation ceiling, and the marginal value of a sample. Section 4 separates coverage from selection, explains the identifiability gap between them, and measures it on both an independent-draw log [2] and a dependent-draw log [20]. Section 5 decomposes within- and between-problem correlation, gives the compute-allocation rule, an estimator for ρ, and a summary table.

2 Test-time sampling is cluster sampling

The entire argument rests on one reframing: a problem and its repeated attempts form a cluster, not a fresh draw each time. This section makes the correspondence precise and fences exactly where the claims that follow are exact.

Fix a prompt q and draw n responses $o_1, \ldots, o_n$ from one model at a fixed decoding configuration (the same sampling settings each time). A verifier (an automatic checker that marks each answer right or wrong) scores each, giving binary success indicators

$$Y_i = \mathbf{1}\{\, o_i \text{ is correct for } q \,\} \in \{0,1\}, \qquad i = 1, \ldots, n. \tag{2}$$

Write $s = \mathbb{P}[Y_i = 1]$ for the per-attempt success probability and $K = \sum_{i=1}^{n} Y_i$ for the number correct. The three

headline quantities of test-time scaling are functions of $(Y_1, \ldots, Y_n)$:

$$\underbrace{\text{pass@}n = \mathbb{P}[K \geq 1]}_{\text{coverage}}, \qquad \underbrace{\hat{p} = K/n}_{\text{success fraction}}, \qquad \underbrace{\mathbf{1}\{K > n/2\}}_{\text{majority vote}}. \tag{3}$$

Best-of-n (drawing n answers and returning the one a learned reward model scores highest) [21], weighted voting, and self-consistency (returning the answer the samples agree on most often) are all read off the sampled distribution and so are functions of the same draws.

The independence baseline assumes $Y_1, \ldots, Y_n$ are i.i.d. (independent and identically distributed) Bernoulli(s) draws, each a coin that lands correct with probability s. Then pass@$n = 1-(1-s)^n$, $\mathrm{Var}(\hat{p}) = s(1-s)/n$, and, for $s > \frac{1}{2}$, the majority vote is correct with probability tending to 1 as $n \to \infty$ [9]. The analysis keeps the marginal s and replaces independence with exchangeability.

Exchangeable attempts. The attempts to one problem are exchangeable: relabeling them does not change their joint distribution, so only how many succeed matters, not which ones. By de Finetti's theorem [22], an infinitely exchangeable binary sequence (one a fresh session can in principle extend to arbitrarily many attempts) is a mixture of i.i.d. sequences,

$$Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta), \qquad \theta \sim G \text{ on } [0,1], \tag{4}$$

for some mixing distribution G (the spread of these hidden success rates across problems) with mean $\mathbb{E}[\theta] = s$. In words, the model behaves as if each fresh session first draws a hidden success rate θ for the problem, then every attempt in that session is an independent θ-weighted coin: a good session lands in a strong reasoning basin (large θ), a weak one in a poor basin (small θ). The single number that summarizes the dependence is the *intraclass correlation* ρ, how strongly two attempts on the same problem move together, and the standard measure for clustered binary data,

$$\rho \;=\; \mathrm{Corr}(Y_i, Y_j) \;=\; \frac{\mathrm{Var}(\theta)}{s(1-s)} \;\in\; [0,1], \qquad i \neq j. \tag{5}$$

The representation forces $\mathrm{Var}(\theta) \geq 0$, so $\rho \geq 0$: exchangeable attempts are non-negatively correlated. Independence is $\rho = 0$ (a degenerate G at s); $\rho = 1$ is total collapse, where a session is all correct or all wrong together. Equation (5) is the bridge from the spread of the latent rate θ to the intraclass correlation that drives the design effect. This ρ is a correlation of *correctness*; the concentration of the answer *strings*, which governs selection, is a separate quantity, introduced in Section 4.2. Figure 3 states the correspondence.

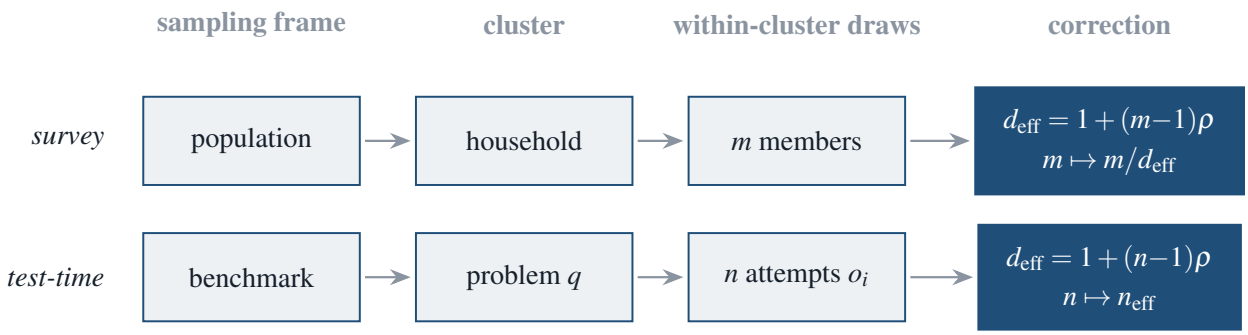

Figure 3: The survey-to-test-time correspondence. *Top:* a survey, where a household of m members maps through $d_{\text{eff}} = 1+(m-1)\rho$. *Bottom:* test-time sampling, where a problem's n attempts map through $d_{\text{eff}} = 1+(n-1)\rho$ to n_{eff}.

2.1 Scope of the exact claims

A canonical claim is only as strong as the assumptions it names, and the main one here is weaker than it looks. The load-bearing assumption is *exchangeability*: the attempts to a problem are taken order-free, which gives them a common success rate s and a common pairwise correlation ρ and makes $d_{\text{eff}} = 1 + (n-1)\rho$ and the ceiling $1/\rho$ exact. Little of this rests on the correlation being the *same* for every pair. The variance identity of Proposition 1 needs only the common rate s: for any pattern of pairwise correlations it holds with ρ read as the *mean* pairwise correlation $\bar{\rho} = \frac{1}{n(n-1)} \sum_{i \neq j} \mathrm{Corr}(Y_i, Y_j)$, so $d_{\text{eff}} = 1 + (n-1)\bar{\rho}$ and the ceiling is $1/\bar{\rho}$. Real sampling departs from equicorrelation (attempts that share a long prefix are more alike than attempts that branch early); then $1/\rho$ is simply read as $1/\bar{\rho}$, the average the estimator of Section 5.3 already returns. What the ceiling truly requires is only that the correlation be positive and not vanish, that is, that sampling diversity stay bounded: the de Finetti mixture forces $\mathrm{Var}(\theta) \geq 0$, hence $\rho \geq 0$, and infinite exchangeability, the one further assumption, carries every $n \to \infty$ limit below. Equicorrelation is thus a convenience that turns an average into a single number, not a crutch the ceiling stands on.

Under these assumptions the design effect in Section 3 is *exact* for one estimand: the success fraction $\hat{p} = K/n$, which depends on (Y_i) only through their sum and so estimates the per-problem rate s or, pooled, the benchmark mean. It governs how precisely a sampling budget pins down accuracy, the estimation use of test-time sampling. Two other quantities read the same draws but are not this mean and are treated on their own terms. Coverage, pass@$n = \mathbb{P}[K \geq 1]$, depends on the full mixing distribution G, and Section 4.1 treats it directly through Equation (4). Selection (self-consistency, best-of-n) returns the most-favored answer, the mode of the categorical answer distribution rather than a sum of the Y_i, and Section 4.2 gives it its own ceiling. The three are kept apart on purpose, because their different dependence on the same draws is the point here, and it is what lets one budget buy different amounts of each. The verifier is assumed accurate; an imperfect verifier adds a second, independent ceiling on usable accuracy, one not folded into ρ here.

3 The effective number of samples

Everything downstream, the ceiling and the budget rule, follows from a single quantity: the variance of the number correct. This section computes it and reads off the effective number of samples.

Proposition 1 (Design effect of test-time sampling)**.** *Let $Y_1, \ldots, Y_n$ be exchangeable with $\mathbb{P}[Y_i = 1] = s$ and $\mathrm{Corr}(Y_i, Y_j) = \rho$ for $i \neq j$. Then the count $K = \sum_i Y_i$ has*

$$\mathbb{E}[K] = ns, \qquad \mathrm{Var}(K) = ns(1-s)\big[\,1 + (n-1)\rho\,\big], \tag{6}$$

and the success fraction $\hat{p} = K/n$ has $\mathrm{Var}(\hat{p}) = s(1-s)/n_{\text{eff}}$ with the effective number of samples

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\rho}. \tag{7}$$

Proof. Linearity gives $\mathbb{E}[K] = ns$. For the variance, $\mathrm{Var}(K) = \sum_i \mathrm{Var}(Y_i) + \sum_{i \neq j} \mathrm{Cov}(Y_i, Y_j)$. Each $\mathrm{Var}(Y_i) = s(1-s)$ and each of the $n(n-1)$ covariances equals $\rho\, s(1-s)$, so $\mathrm{Var}(K) = ns(1-s) + n(n-1)\rho\, s(1-s) = ns(1-s)[1 + (n-1)\rho]$. Dividing by n^2 gives $\mathrm{Var}(\hat{p}) = s(1-s)[1 + (n-1)\rho]/n = s(1-s)/n_{\text{eff}}$. □

The bracket $d_{\text{eff}} = 1 + (n-1)\rho$ is the design effect: the factor by which correlation inflates the variance of the count relative to n independent draws [10, 11]. Equivalently, the $n \times n$ covariance of the attempts is a constant diagonal plus a single rank-one term, $\Sigma = s(1-s)\big[(1-\rho)\mathbf{I} + \rho\, \mathbf{1}\mathbf{1}^{\top}\big]$, so the variance of the count collapses to the scalar d_{eff} without ever forming that matrix: the same structured-operator economy that lets large numerical systems be solved matrix-free [23]. The estimator $\hat{p}$ behaves as if it were built from n_{eff} independent samples; Appendix B confirms the variance identity (6) against Monte Carlo simulation.

3.1 The correlation ceiling

This is the result this work is named for, and its single most important fact: the effective number of samples does not grow without bound. Equation (7) has a finite limit.

Corollary 1 (Correlation ceiling)**.** *If $\rho > 0$, then n_{eff} increases in n to the finite limit*

$$\lim_{n\to\infty} n_{\text{eff}} = \frac{1}{\rho}, \tag{8}$$

and n_{eff} reaches half of this ceiling at $n = (1-\rho)/\rho \approx 1/\rho$.

Proof. Write $n_{\text{eff}} = n/[\,1-\rho+n\rho\,] \to 1/\rho$ as $n \to \infty$. Setting $n_{\text{eff}} = 1/(2\rho)$ gives $2\rho n = 1+(n-1)\rho$, i.e. $\rho n = 1-\rho$, so $n = (1-\rho)/\rho$. □

This is the central fact. A correlated cluster with intraclass correlation ρ is worth at most $1/\rho$ independent draws, however large the budget (exactly under the equicorrelation of Section 2.1, and as a limiting average otherwise); and it gets halfway there by $n \approx 1/\rho$. Figure 4 plots the ceiling. A model with $\rho = 0.1$ caps out near ten effective samples: the thousandth draw is almost worthless. The ceiling is also why the marginal value of a sample collapses.

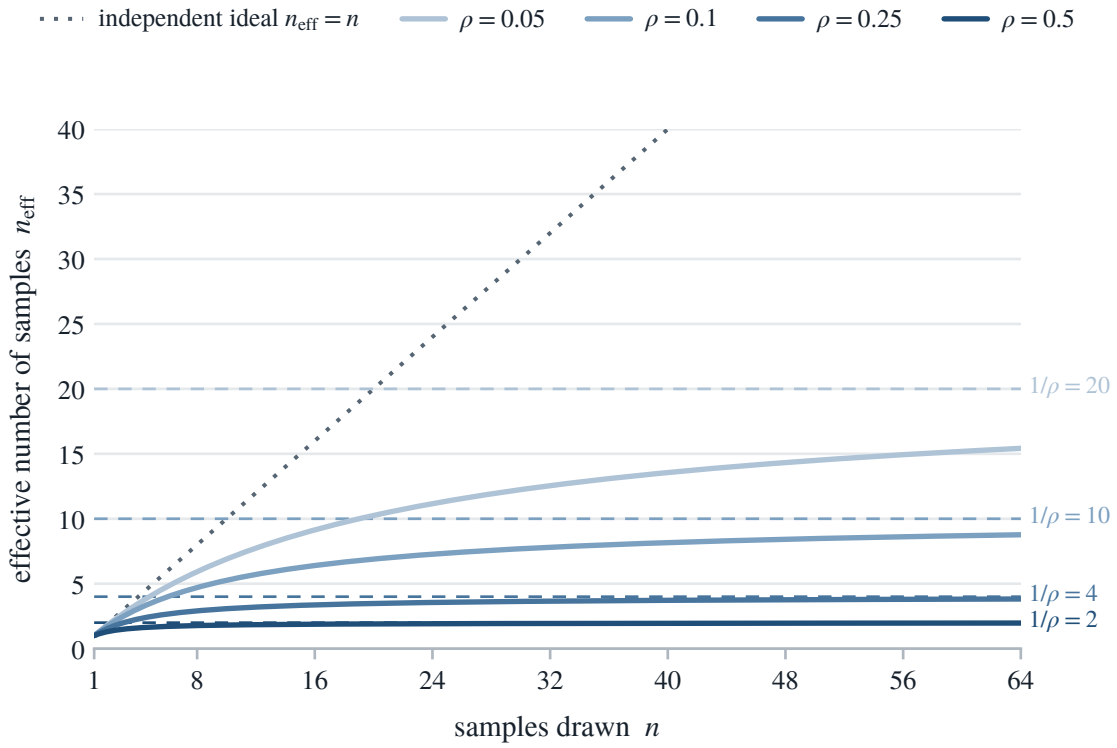

Figure 4: The correlation ceiling. The effective number of samples $n_{\text{eff}} = n/[1+(n-1)\rho]$ against the number drawn, for four correlation levels, each saturating at its ceiling $1/\rho$ (dashed); the independent ideal $n_{\text{eff}} = n$ is the diagonal.

Corollary 2 (Value of the n-th sample)**.** *The marginal effective sample contributed by the n-th draw is*

$$\frac{d\,n_{\text{eff}}}{dn} = \frac{1-\rho}{\big[\,1+(n-1)\rho\,\big]^2} \sim \frac{1-\rho}{\rho^2 n^2} \quad (n\to\infty), \tag{9}$$

so the worth of an added sample decays quadratically and is negligible once $n \gg 1/\rho$.

The marginal value starts at $d\,n_{\text{eff}}/dn = 1-\rho$ at $n=1$ (the second draw already adds only $(1-\rho)/(1+\rho)$ effective samples) and then falls off as $1/(\rho n)^2$. Figure 5 shows the collapse and marks the break-even point $n \approx 1/\rho$ at which the curve has already given up most of its value. This is a budget rule, derived in Section 5.2: spending past $1/\rho$ samples buys redundancy, not signal.

A worked reading makes the decay concrete. At $\rho = 0.1$ the first draw is worth a full independent sample; by Corollary 2 the tenth is worth $(1-\rho)/[1+9\rho]^2 = 0.9/1.9^2 \approx 0.25$, and the hundredth only $0.9/10.9^2 \approx 0.008$. A thousand-sample run therefore carries the estimation information of roughly its first ten draws, and the rest refine the estimate by almost nothing. The decay is quadratic, so each tenfold increase in budget past the ceiling returns about a hundredth as much, which is why the curves of Figure 5 fall to the axis so quickly.

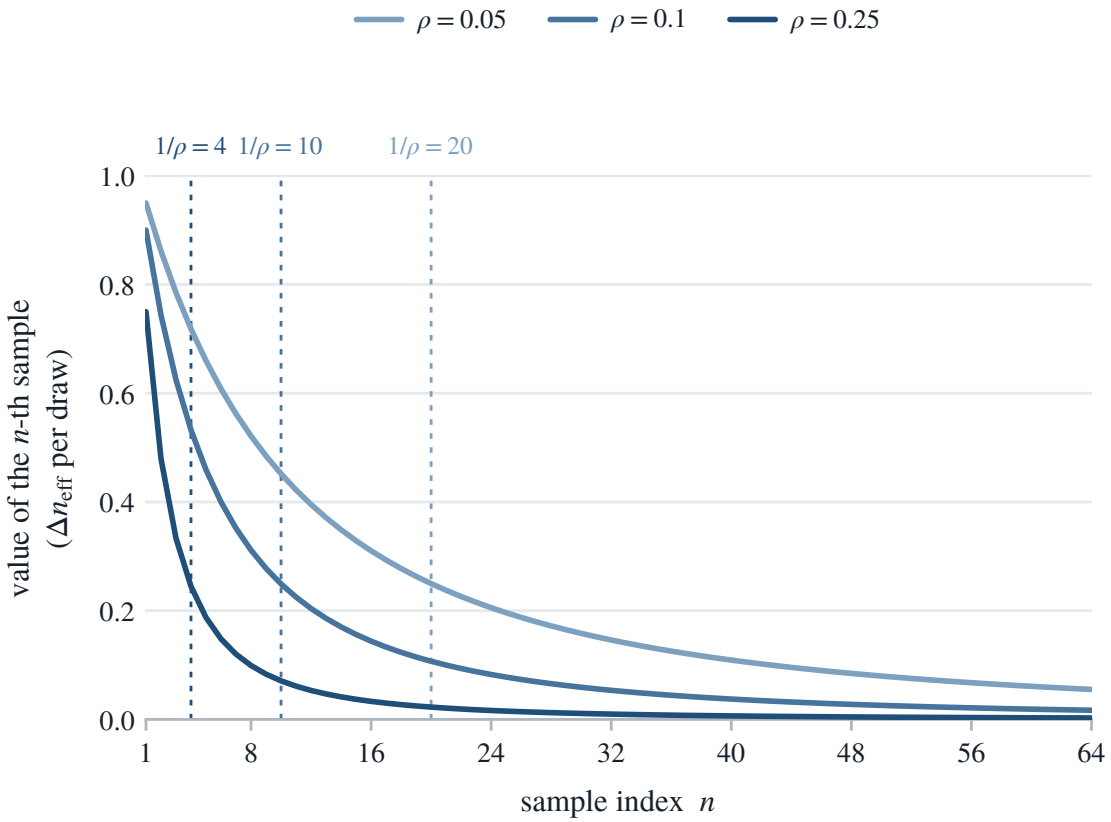

Figure 5: The value of the n-th sample, $d\,n_{\text{eff}}/dn$ from Corollary 2, against the sample index. Break-even points $n \approx 1/\rho$ marked dotted (one per ρ).

4 Coverage rises, selection saturates

The most-cited puzzle in test-time scaling is that one curve keeps rising while another stalls, and no single mechanism connects them. This section gives the mechanism: coverage and selection read the same samples for different things, so they obey different ceilings. Coverage is capped only by what the model could ever produce; selection by whether its most common answer is correct. The gap between the two is exactly the solvable problems whose correct answer the vote fails to return.

Why selection matters at all bears stating, since coverage looks like the whole story. Coverage can be cashed

only through a verifier that certifies a correct answer, and a perfect, general verifier is the rare exception rather than the rule. A few domains supply one: code against a test suite, a proof against a proof assistant, a numeric answer against a key. Most do not, and the verifiers that do exist score a proxy rather than truth, a finite test suite or a learned reward model that a wrong answer can still pass. Wherever no sound verifier is available, which is the typical deployment, a single answer can be returned only by selection, so the modal ceiling of Section 4.2, not coverage, fixes the accuracy a system delivers. Coverage measures what the model could reach given a verifier it will not have; selection measures what ships.

4.1 Coverage and the correlation tax

Coverage is the optimistic half of the split, the curve that keeps climbing; this subsection prices what correlation quietly subtracts from it. Coverage asks only whether some sample is correct, with no need to know which one. Under the mixture (4),

$$\text{pass@}n = 1 - \mathbb{P}[K = 0] = 1 - \mathbb{E}_\theta\left[(1-\theta)^n\right]. \tag{10}$$

Coverage can only rise as the budget grows, and outside a hard core of unreachable problems it rises all the way.

Proposition 2 (Coverage rises without a within-problem ceiling)**.** *Under the mixture* (4)*,* pass@n *is non-decreasing in n, strictly increasing whenever G places mass on $(0,1)$, and converges to $1-\pi_0$, where $\pi_0 = G(\{0\})$ is the fraction of problems the model never solves. For a fixed problem with reachability $p_q(c_q) = \pi > 0$, coverage $1-(1-\pi)^n$ strictly increases to* 1*.*

Proof. For each $\theta \in [0,1]$ the map $n \mapsto (1-\theta)^n$ is non-increasing, strictly for $\theta \in (0,1)$. Taking the mixture expectation, $\mathbb{P}[K=0] = \mathbb{E}_\theta[(1-\theta)^n]$ is non-increasing in n, strictly if G charges $(0,1)$, so pass@$n = 1-\mathbb{P}[K=0]$ is non-decreasing, strictly so. Since $(1-\theta)^n \to \mathbf{1}\{\theta = 0\}$ pointwise, dominated convergence gives $\mathbb{P}[K=0] \to \pi_0$, hence pass@$n \to 1-\pi_0$. The fixed-problem case is the point mass $\theta = \pi$. □

Correlation always taxes coverage relative to the independent reading.

Proposition 3 (Correlation tax on coverage)**.** *For exchangeable attempts with per-attempt success s,*

$$\text{pass@}n \;\le\; 1-(1-s)^n, \tag{11}$$

with equality if and only if $\rho = 0$. The gap grows with the dispersion of θ: a mean-preserving spread of the difficulty distribution (more very-easy and very-hard problems at the same average) can only widen it.

Proof. The map $\theta \mapsto (1-\theta)^n$ is strictly convex on $[0,1]$ for $n \ge 2$. By Jensen's inequality (the average of a convex function is at least the function of the average), $\mathbb{E}_\theta[(1-\theta)^n] \ge (1-\mathbb{E}[\theta])^n = (1-s)^n$, with equality iff θ is degenerate, i.e. $\rho = 0$. Negating gives Equation (11). □

So the textbook curve $1-(1-s)^n$ is an upper bound, not a prediction: real coverage runs below it whenever samples are correlated. The shape of the shortfall is governed by the lower tail of G, the problems the model rarely solves.

Proposition 4 (Power-law coverage from heterogeneity)**.** *Let $\theta \sim \text{Beta}(\alpha,\beta)$ (a flexible family of success-rate distributions on the unit interval, standing in for the spread of problem difficulty), so $s = \alpha/(\alpha+\beta)$ and $\rho = 1/(\alpha+\beta+1)$. Then the miss rate is*

$$\mathbb{P}[K=0] = \frac{B(\alpha,\beta+n)}{B(\alpha,\beta)} \;\sim\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\, n^{-\alpha} \qquad (n \to \infty), \tag{12}$$

so coverage approaches its limit as a power law $n^{-\alpha}$, not the exponential $(1-s)^n$.

Proof. With $\theta \sim \text{Beta}(\alpha,\beta)$, $\mathbb{E}[(1-\theta)^n] = B(\alpha,\beta+n)/B(\alpha,\beta)$ is the standard beta-binomial (a binomial whose success rate is itself drawn from the Beta) probability of zero successes. Writing it as $\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \cdot \frac{\Gamma(\beta+n)}{\Gamma(\alpha+\beta+n)}$ and using $\Gamma(\beta+n)/\Gamma(\alpha+\beta+n) \sim n^{-\alpha}$ (Stirling) gives the tail. The intraclass correlation of a beta-binomial is $\rho = 1/(\alpha+\beta+1)$. □

Proposition 4 is the difficulty-heterogeneity story of Schaeffer et al. [17] and Kazdan et al. [18], recovered here as the coverage face of the same model for a Beta difficulty prior: a heavy lower tail (small α) yields the slowly rising, log-linear coverage that Brown et al. [2] fit empirically, with the exponent set by the lower tail of G. Figure 6 contrasts the exponential and power-law regimes. The overdispersed binomial behind it (a count with more spread than independent draws would give) is classical [24]. If G additionally places an atom π_0 at $\theta = 0$, a hard core of attempts the model never makes correct, then pass@$n \to 1-\pi_0 < 1$ and no budget crosses it; this is mode collapse in its starkest form. The atom is a ceiling of model capability rather than of sampling: those problems lie outside the model's reach, a generalization limit that no sampling budget repairs [25].

The practical force of the power law is how slowly it pays. An exponential miss rate $(1-s)^n$ clears any target in a handful of samples, each one cutting the miss rate by a constant factor; a power-law miss rate $n^{-\alpha}$ does not. Halving an exponential miss rate costs a fixed number of further samples; halving a power law costs a fixed *multiplicative* factor $2^{1/\alpha}$, so for a heavy lower tail (small α) each successive halving costs as much sampling as everything before it combined. Coverage keeps rising, but with sharply diminishing speed: the slow, log-linear climb that the released logs show over four orders of magnitude. Figure 6 plots both regimes on log–log axes, where the exponential falls off a cliff and the power law settles onto a straight line of slope $-\alpha$.

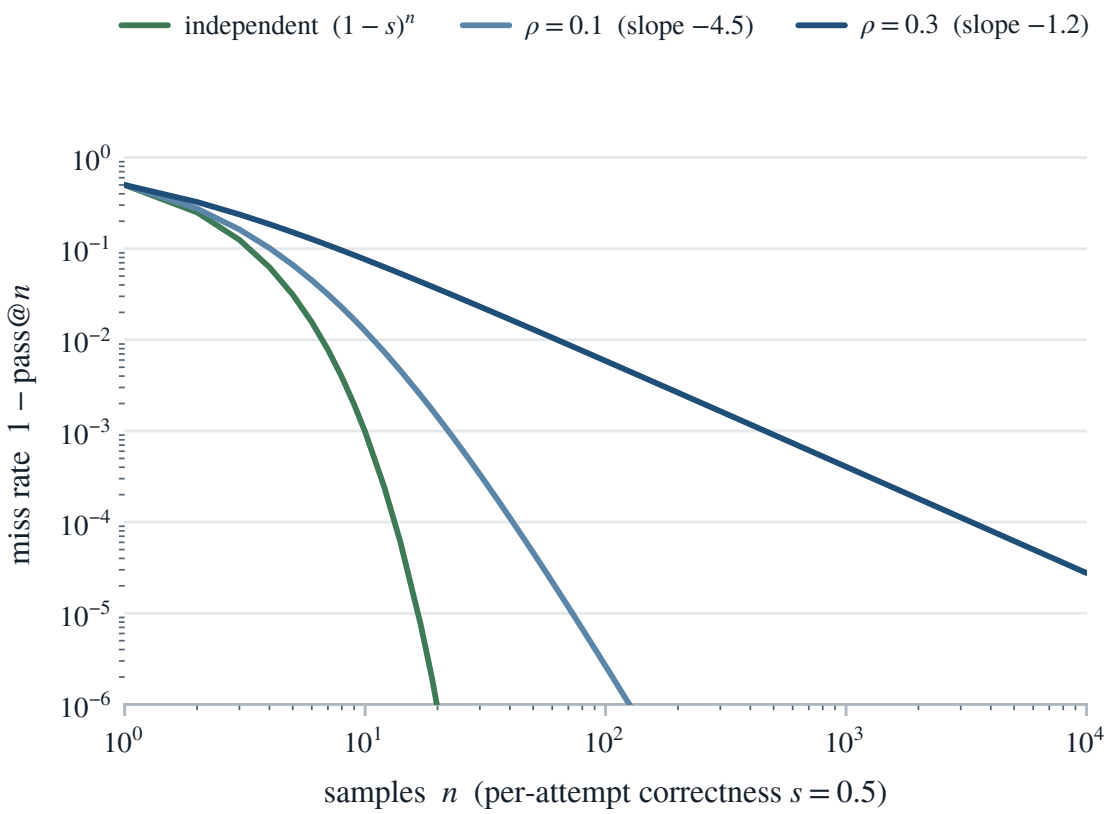

Figure 6: Exponential versus power-law coverage. The miss rate $1 - \text{pass@}n$ on log–log axes: the independent exponential against the power law $n^{-\alpha}$ of Proposition 4. Here $s = 0.5$ and $\rho \in \{0.1, 0.3\}$.

4.2 Selection and the modal answer

Selection is where the visible gains stop, and the cause is not correlation but the shape of the answer distribution. To return an answer without a verifier, a method keeps the most common answer (self-consistency [1]) or the highest-scored one (reward-model best-of-n); either way it reads the whole distribution over answer strings, not just the per-attempt correctness rate. Fix a problem q and let the model induce a distribution $p_q(a)$ over candidate answers a, with correct answer c_q and a unique most common answer, the *mode* $a_q^\star = \arg\max_a p_q(a)$. A plurality vote among n attempts returns the empirical mode.

Proposition 5 (Selection ceiling)**.** *As $n \to \infty$ the plurality answer converges almost surely to the mode $a_q^\star$, so self-consistency accuracy on a fixed problem tends to $\mathbf{1}\{a_q^\star = c_q\}$. Averaged over a benchmark, self-consistency accuracy converges to the* modal-hit rate

$$\pi_{\text{mode}} \;=\; \mathbb{P}_q\big[a_q^\star = c_q\big], \tag{13}$$

a hard ceiling, the modal ceiling*, with no dependence on the sample budget n.*

Proof. For a fixed problem the empirical answer frequencies converge to p_q almost surely (the law of large numbers applied to the multinomial counts), and arg max is continuous wherever the maximizer is unique, so the empirical mode converges to $a_q^\star$ almost surely. Hence $\mathbf{1}\{\text{plurality correct}\} \to \mathbf{1}\{a_q^\star = c_q\}$, and averaging over problems gives the limit (13), which is constant in n. □

The ceiling is a wall, not a slowdown: once the budget reveals the mode, more samples cannot move the vote, and on the problems where the mode is wrong they move it the wrong way.

Corollary 3 (Anti-scaling of selection)**.** *On any problem whose mode is incorrect ($a_q^\star \neq c_q$) yet whose correct answer is reachable ($p_q(c_q) > 0$), self-consistency accuracy falls to* 0 *while coverage rises to* 1 *as $n \to \infty$: more sampling makes selection worse and coverage better on the same problem.*

Proof. By Proposition 5 the plurality converges to $a_q^\star \neq c_q$, so $\mathbf{1}\{\text{plurality correct}\} \to 0$, while $\text{pass@}n = 1-(1-p_q(c_q))^n \to 1$ because $p_q(c_q) > 0$. □

Anti-scaling is the mechanism behind the limited and sometimes non-monotone returns that self-consistency and reward-model selection show in practice [26, 27]: extra samples sharpen a confident wrong answer. How fast a problem reaches its ceiling is set by the concentration of p_q, summarized by the *effective number of answers* $1/\sum_a p_q(a)^2$; the ceiling *height* π_{mode} is set by whether the mode it converges to happens to be correct. Figure 7 shows both outcomes: coverage rises to one whether or not the mode is correct, while selection rises to one only when the mode is correct and otherwise falls to zero. Self-consistency and plurality voting meet this modal ceiling exactly. Best-of-n with a learned reward model is bounded instead by how well that scorer ranks answers: a perfect verifier lifts it to coverage, a frequency-like scorer collapses it to the modal ceiling, and real reward models fall between.

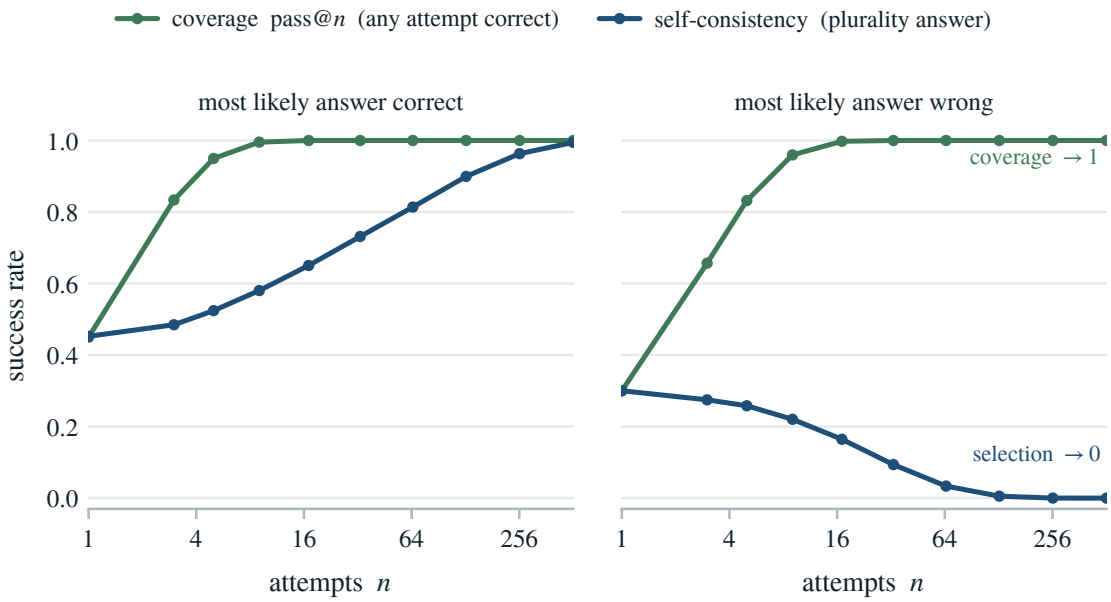

Figure 7: Anti-scaling of selection. Two problems sampled to n attempts each: coverage (any attempt correct) and self-consistency (the plurality answer) against n. Left: a problem whose most common answer is correct. Right: a problem whose most common answer is wrong.

The familiar majority-vote reading is the binary special case, and it is a strict lower bound. Collapse the answers to two, the correct one against a single modal error, and let θ be the latent per-attempt correctness of Equation (4); then plurality is majority, $K/n \to \theta$ almost surely, and the vote is correct in the limit exactly when $\theta > \frac{1}{2}$.

Corollary 4 (Condorcet jury, correlated)**.** *If the difficulty distribution places positive mass on problems the model gets wrong more often than right ($\mathbb{P}[\theta < \frac{1}{2}] > 0$), majority-vote accuracy converges to $\mathbb{P}[\theta >$*

$\frac{1}{2}] < 1$, *below the value* 1 *that independent jurors reach when* $s > \frac{1}{2}$ *[12, 13]; the normal approximation* $\theta \approx \mathcal{N}(s, \rho\, s(1-s))$ *gives* $\mathbb{P}[\theta > \frac{1}{2}] \approx \Phi\big((s-\frac{1}{2})/\sqrt{\rho\, s(1-s)}\big)$. *This never exceeds the plurality ceiling,* $\mathbb{P}[\theta > \frac{1}{2}] \le \pi_{\text{mode}}$, *because* $\theta > \frac{1}{2}$ *forces every wrong answer below* $\frac{1}{2}$ *and so below the correct one.*

The two bounds part company on real logs, and the distance between them is the point. On the dependent-draw log of Section 4.3 the majority bound reads $\mathbb{P}[\theta > \frac{1}{2}] = 0.20$ while the plurality ceiling is $\pi_{\text{mode}} = 0.45$: more than half of the selectable accuracy comes from problems the model answers correctly less than half the time, where the correct answer is still the single most common one because the errors scatter. Figure 8 shows the binary plateau and that simulated accuracy lands on it.

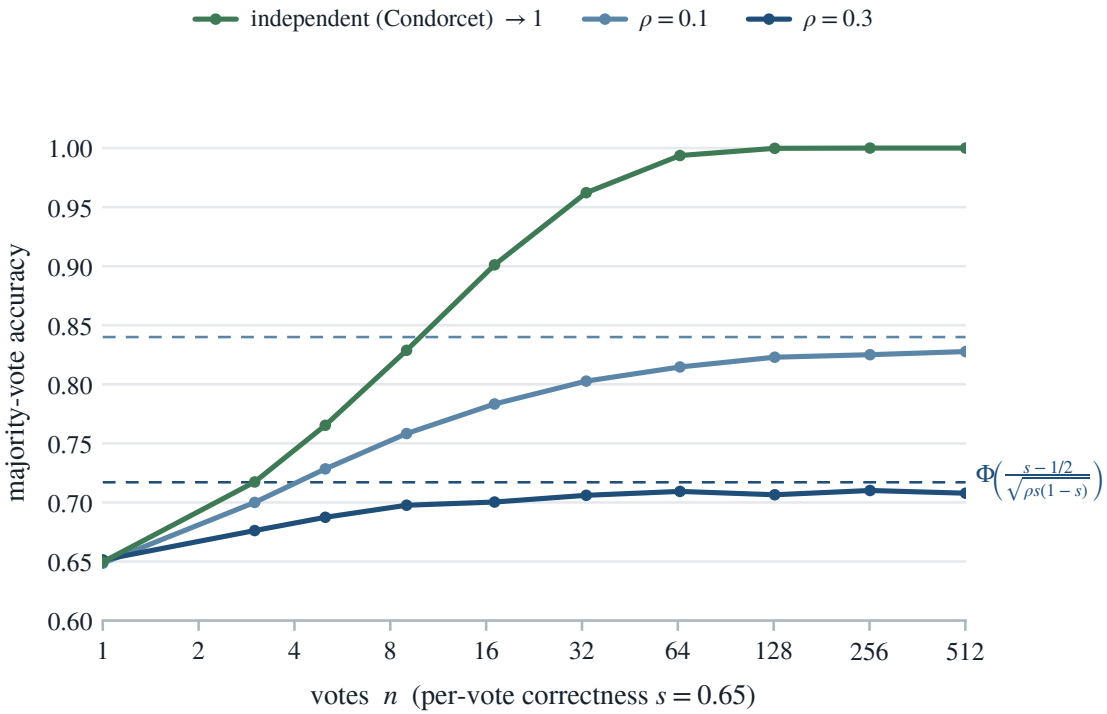

Figure 8: The binary (majority-vote) special case of the selection ceiling. Majority-vote accuracy against n for several ρ: the independent case ($\rho \to 0$, top) and correlated cases plateauing at $\mathbb{P}[\theta > \frac{1}{2}]$ (Corollary 4), with its normal approximation dashed. Markers: simulation; $s = 0.65$.

4.3 The coverage–selection gap, measured

Theory has predicted a gap; this is where it is measured on real sampling logs. Put the two together. Coverage (Proposition 3) climbs toward what the model can reach, $1 - \pi_0$, slowly. Selection (Proposition 5) saturates at the modal-hit rate π_{mode}. The two diverge, and the divergence is the identifiability gap: the problems whose correct answer is in the pool but is not its most common answer. The split shows up in two released logs that probe the two effects separately: an independent-draw log for the between-problem difficulty spread, and a dependent-draw log for the within-problem answer collapse.

The between-problem term, on independent draws. Brown et al. [2] sampled up to 10^4 solutions per problem on GSM8K [21] (grade-school math word problems) and MATH [28] (harder competition mathematics) and recorded the correctness of each. Their attempts are drawn independently, so the within-problem correlation is zero by construction; what the logs expose is the *between-problem* difficulty spread ρ_b and its consequences: the $\rho_w = 0$ face of the model. The analysis here estimates, per configuration, the

difficulty mean s, the intraclass correlation $\hat{\rho} = \mathrm{Var}(\theta)/[s(1-s)]$, coverage by the unbiased estimator of Chen et al. [9], and self-consistency by plurality vote of the extracted answers (validated against the logs' own correctness labels). The self-consistency plateau is reported only where the numeric extractor reproduces those labels; MATH answers are boxed expressions a numeric extractor cannot parse, so MATH shows coverage and $\hat{\rho}$ alone. Table 1 reports them.

Table 1: The between-problem difficulty correlation $\hat{\rho}_b$ and the coverage–selection gap, from the released logs of Brown et al. [2] (10^4 samples per problem). The $\hat{\rho}_b$ 95% CI is a problem-level clustered bootstrap (10^4 resamples). Self-consistency is n/a on MATH (boxed answers a numeric extractor cannot parse).

Benchmark	**Model**	s	$\hat{\rho}_b$	$\hat{\rho}_b$ 95% CI	$1/\hat{\rho}_b$	coverage@10^4	self-cons.
GSM8K	Llama-3-8B-Instruct	0.77	0.47	[0.40, 0.54]	2.1	1.00	0.87
GSM8K	Llama-3-70B-Instruct	0.93	0.41	[0.24, 0.54]	2.5	1.00	0.97
MATH	Llama-3-8B-Instruct	0.27	0.48	[0.41, 0.55]	2.1	0.98	n/a
MATH	Llama-3-70B-Instruct	0.48	0.61	[0.54, 0.66]	1.6	0.98	n/a

Two readings stand out. First, the difficulty correlation is large and stable, $\hat{\rho} \approx 0.4$–0.6 across models and benchmarks, so by Equation (7) the 10^4 samples drawn for each problem carry the benchmark-mean information of only $n_{\mathrm{eff}} \approx 2$ independent samples: for estimating benchmark accuracy, ten thousand samples of one problem are worth about two independent ones. Second, coverage and selection diverge in the direction the split predicts. Figure 9 shows the GSM8K, Llama-3-8B-Instruct curves: coverage reaches 1.00 while self-consistency plateaus at 0.87. For roughly an eighth of problems the correct answer is somewhere in the pool but the vote does not return it even at the plateau: the usable-signal gap, on real data.

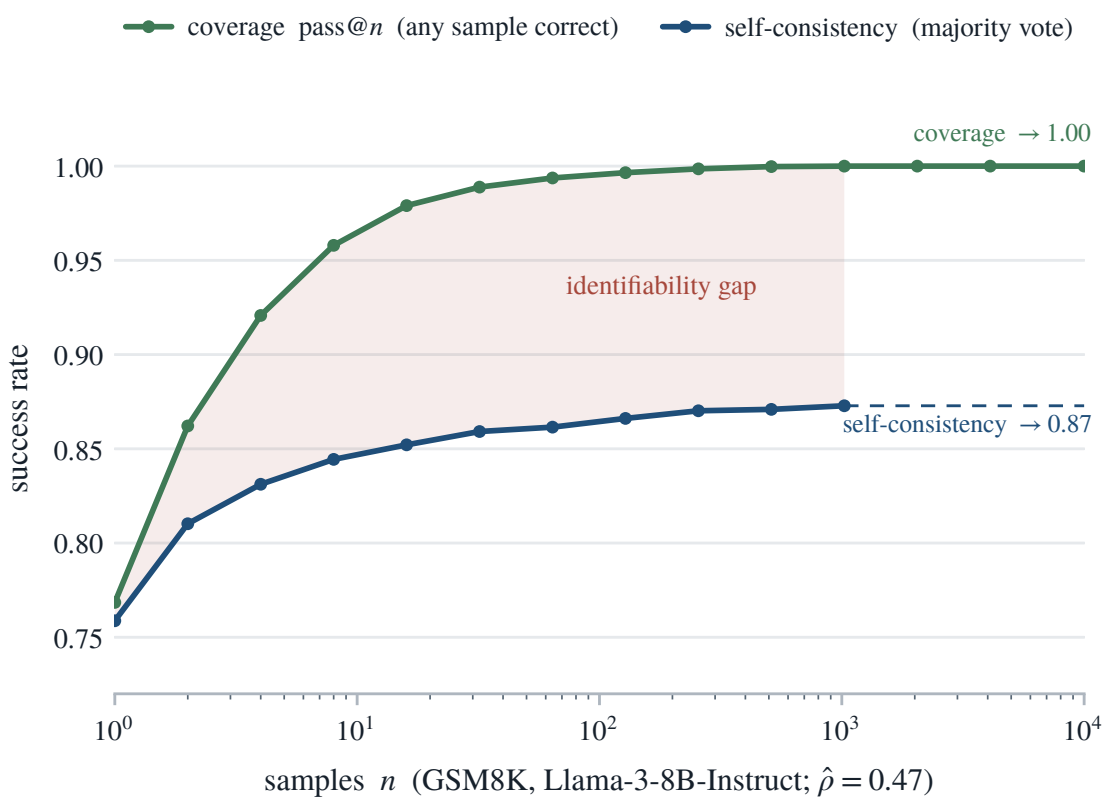

Figure 9: The coverage–selection gap on real logs [2]: GSM8K, Llama-3-8B-Instruct, up to 10^4 samples per problem. Coverage rises to 1.00; self-consistency plateaus at 0.87; the shaded band is the usable-signal gap.

The reading is consistent across the repeated-sampling literature. Chen et al. [26] find voting performance non-monotone in the number of calls because a benchmark mixes easy and hard problems, the anti-scaling of Corollary 3; Wang et al. [27] show that prompt-diversity interventions, which spread the answer distribution, are what move best-of-n; and Kirk et al. [29] document that post-training (the fine-tuning applied after pretraining) sharpens the modal answer. Each is a face of one split: selection is held at its modal ceiling, coverage is not.

The selection ceiling, on raw answers. The modal-hit rate π_{mode} needs the answer strings, which the correctness-only logs above do not record; a log of raw completions supplies them. The best-of-n release of Beeching et al. [20] samples each of the 500 MATH-500 problems 256 times from one model (Llama-3.2-1B-Instruct) at a fixed decoding configuration (temperature 0.8, top-p 1.0), recording every raw completion. Across the 500 problems the 256 answers carry a median of only about thirteen distinct values (an effective answer count $1/\sum_a p_a^2$), not 256: the answer distribution p_q is sharply concentrated. Figure 10 measures the consequence: coverage (any attempt correct) climbs to 0.88 while self-consistency (the plurality answer) plateaus at $\pi_{\text{mode}} = 0.45$ by about $n = 64$ and moves little after. Correctness is graded with the same verifier that produced the dataset's labels (`math-verify`), reproducing its reported single-sample accuracy (≈ 0.27) to about a point. The binary majority bound on the same log reads only $\mathbb{P}[\theta > \frac{1}{2}] = 0.20$: half the realized selection accuracy comes from problems solved less than half the time whose scattered errors leave the correct answer as the single most common one. The plateau is the modal ceiling of Proposition 5, on real attempts: the correct answer reaches the pool for nearly nine problems in ten, but the most common answer is correct for only four and a half.

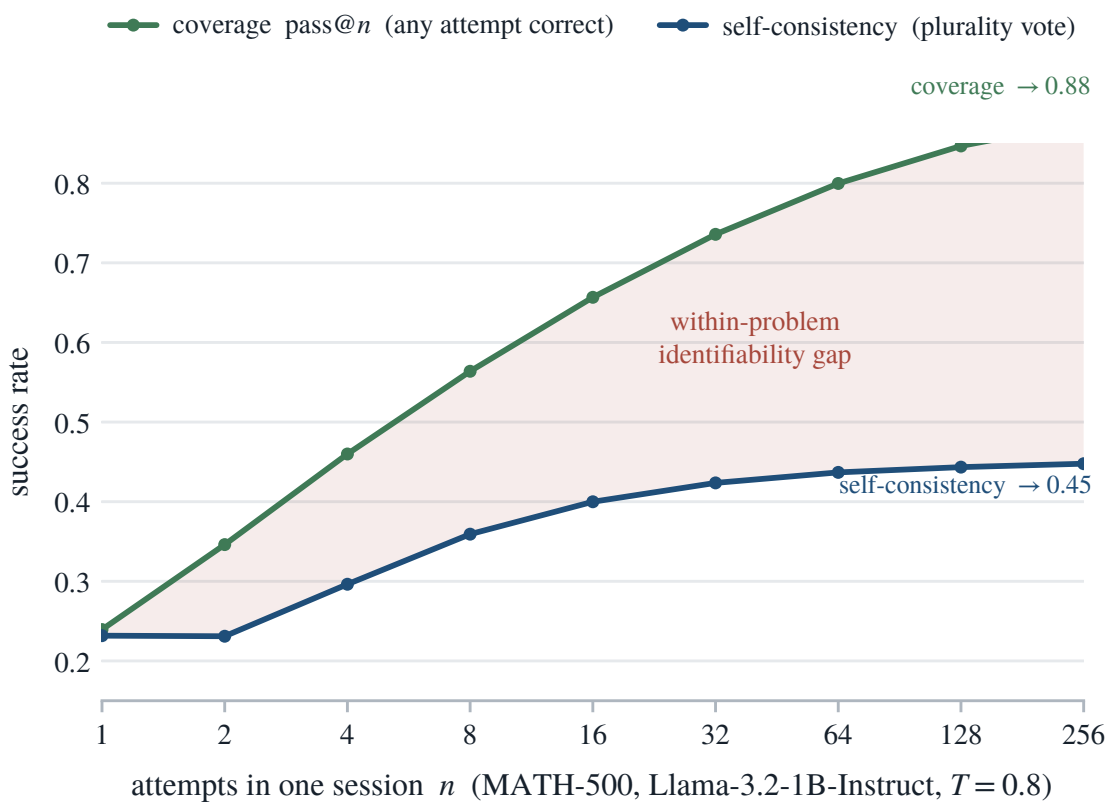

Figure 10: The within-problem coverage–selection gap on a dependent-draw log [20]: MATH-500, Llama-3.2-1B-Instruct, 256 attempts per problem in one session (averaged over five sessions). Coverage rises to 0.88; self-consistency plateaus at 0.45; the shaded band is the within-problem usable-signal gap.

The five sessions also test the decomposition itself. Estimating ρ_w as the run-to-run dispersion of a problem's success rate across the five independent sessions, corrected for the binomial noise of a finite session, gives $\hat{\rho}_w$ near zero ($\hat{\rho}_w \approx 0.0007$, 95% CI $[0.0005, 0.0009]$): re-running a fixed model at a fixed temperature does not move the latent success rate beyond sampling noise, so the de Finetti rate θ is nearly fixed per problem and the same-session pooled correlation is dominated by difficulty. The two-stage identity then holds on real data to within 0.001: the directly pooled same-session correlation is $\hat{\rho} = 0.401$, against $\hat{\rho}_b + (1-\hat{\rho}_b)\hat{\rho}_w = 0.402$ from the separate terms (Equation (14), with this log's own $\hat{\rho}_b = 0.402$). The selection plateau in Figure 10 therefore is not a between-session effect: it is the within-session answer collapse, the concentration of the answer distribution p_q rather than any drift in the success rate, which Proposition 5 prices as the modal-hit rate π_{mode} and Corollary 4 lower-bounds by $\mathbb{P}[\theta > \frac{1}{2}]$.

> Repeated sampling answers two questions that a single accuracy number blurs. *Coverage* asks whether a correct answer is reachable at all, and keeps rising with the budget. *Selection* asks whether the model's most common answer is the correct one, and is capped at the modal-hit rate π_{mode}. Their difference is the *identifiability gap*: the problems a model can reach but not return.

5 Decomposition, allocation, and measurement

A lens earns its place only if it can be measured and acted on. This section splits the correlation into its two sources, turns the ceiling into a compute-allocation rule, and gives a one-line estimator for ρ that runs on any sampling log already in hand.

5.1 Within- and between-problem correlation

Two sources feed the pooled correlation ρ, and separating them is what tells a benchmark mean from a single answer. *Within-problem* correlation ρ_w is run-to-run dependence on a fixed problem: the dispersion of the latent rate θ across independent sessions in Equation (5). Under standard independent decoding this is zero by construction, each attempt an independent draw given the prompt, and Section 4.3 measures it at $\hat{\rho}_w \approx 0.0007$. *Between-problem* heterogeneity ρ_b is the spread of difficulty across a benchmark: two attempts to a randomly drawn problem move together merely because they share that problem's difficulty. For attempts pooled over a benchmark the variance components add, so the pooled intraclass correlation (taking ρ_w common across problems) satisfies

$$\rho \;=\; \rho_b + (1-\rho_b)\,\rho_w, \tag{14}$$

the standard two-stage design effect of clustered sampling [11]: the total same-problem correlation is the between-problem difficulty spread plus the within-problem dependence left after it. With $\rho_w \approx 0$ the pooled ρ is essentially ρ_b, so the correlation ceiling of Section 3 is a statement about *estimating* a benchmark from heterogeneous problems, not about repeated tries on one problem. The slow, power-law coverage of a benchmark is the same ρ_b through Proposition 4, the regime studied by Schaeffer et al. [17]. The selection ceiling is a third axis, the concentration of the answer distribution p_q (Section 4.2), which ρ does not see: difficulty heterogeneity, run-to-run dependence, and answer-mode collapse are three distinct quantities the nominal sample count silently confounds.

5.2 A compute-allocation rule

This is the practical payoff: with a fixed compute budget, the goal decides how to spend it, because the three goals reach their ceilings at different points. Let total inference compute be $C \approx nL$ for n samples of reasoning length L. The levers are more samples (raise n), longer reasoning (raise L, which raises s), more problems (for evaluation), and decorrelation (spread the answer distribution via temperature, nucleus (top-p) sampling, prompt diversity, or mixing models) [7, 27, 30]. The ceilings of Sections 3 and 4.2 (Corollaries 1–2 and Proposition 5) set the stopping point for each:

> To estimate a benchmark mean, $n \approx 1/\rho_b$ samples per problem suffice (about two at the measured ρ_b); spend the rest on more problems. To select an answer without a verifier, sampling helps only until the plurality stabilizes, on the order of the effective answer count $1/\sum_a p_q(a)^2$, and past that it can anti-scale. To cover, that is to find a correct sample for a verifier, there is no within-problem ceiling and more samples keep paying.

The rule reframes the "think longer or sample more" question [3, 4, 31]: sampling pays without limit for coverage, stops early for selection, and stops almost at once for estimation, so the budget should follow the goal rather than a single number. Whether decorrelation actively raises the selection plateau is the lever the protocol below tests.

5.3 Estimating ρ

The ceiling is actionable only if ρ can be read off runs already in hand, and it can. The correlation is measurable from any sampling log that records, for each of M problems, the number correct c_i out of n_i attempts. The standard moment (analysis-of-variance) estimator for clustered binary data [11] compares the between- and within-problem sums of squares of $\hat{p}_i = c_i/n_i$,

$$\hat{\rho} \;=\; \frac{\mathrm{MS}_{\text{between}} - \mathrm{MS}_{\text{within}}}{\mathrm{MS}_{\text{between}} + (n_0 - 1)\,\mathrm{MS}_{\text{within}}}, \tag{15}$$

with n_0 the average cluster size; equivalently, the beta-binomial model of Kazdan et al. [18] has intraclass correlation $\hat{\rho} = 1/(\hat{\alpha} + \hat{\beta} + 1)$ by the identity in Appendix A. Two cautions. On benchmarks with many fully-solved or never-solved problems the within-problem variance is degenerate and Equation (15) can return a small or negative $\hat{\rho}$, which should be clipped at zero; and a pooled $\hat{\rho}$ mixes ρ_b and ρ_w, so it states how much of the nominal n is real only at the level (a single problem or a whole benchmark) at which it was estimated. Table 1 runs the estimator on the logs of Brown et al. [2]; because those attempts are drawn independently, what it recovers is the between-problem term ρ_b, with a clustered (problem-level) bootstrap interval that resamples problems with replacement. The within-problem term needs dependent draws, which Section 4.3 reads off the single-session log of Beeching et al. [20]: the seed-to-seed spread of a problem's success rate is negligible there ($\hat{\rho}_w \approx 0.0007$), so re-decoding one prompt barely moves the latent rate, yet within a session the *answer* distribution still narrows to about thirteen modes and self-consistency plateaus at 0.45. The single-model selection ceiling is therefore set not by run-to-run drift but by the concentration of the answer distribution, the modal-hit rate π_{mode} of Proposition 5. Genuine correlation does cap voting once the voters are distinct models rather than repeated draws of one: across nine distinct judge models the mean pairwise error correlation is $\hat{\rho} \approx 0.39$, so the nine are worth only about two effective votes [14]. For repeated draws of a single model, by contrast, a competition report finds high-temperature sampling already largely decorrelates the errors [16], so what holds selection down there is not run-to-run correlation but the concentration of the answer mode: the handful of effective answers that makes selection saturate so early.

The practical recommendation is one line, meant to be copied into a methods section verbatim:

> Estimate $\hat{\rho}$ from the sampling log via Equation (15), then report the *effective number of samples* $n_{\text{eff}} = n/[1 + (n-1)\hat{\rho}]$ alongside the nominal count n, with the ceiling $1/\hat{\rho}$.

This is the design-effect-corrected count of Kish [10], and it lets a reader see how much of a sampling budget is real.

A pre-registered protocol for the decoding lever. Section 4.3 measures the within-problem ceiling at one decoding configuration: the latent run rate θ is nearly fixed per problem ($\hat{\rho}_w \approx 0$), so the selection plateau is set by the dispersion of the modal answer within a session, not by run-to-run drift. What remains to be measured is the *lever*: whether the decoding choices that decorrelate answers raise the plateau, the prediction the allocation rule rests on. The following test is pre-registered. Fix a benchmark and a model. For each of $M \geq 200$ problems and each of several decoding configurations (a sweep over temperature, nucleus p, and prompt diversity), draw $m \geq 256$ verified attempts; measure the within-session answer-effective count $1/\sum_a p_a^2$, the plurality plateau, and the answer-indicator intraclass correlation. The pre-registered prediction, from Proposition 5, is that configurations with a lower answer correlation have a higher plurality plateau, monotone in the effective answer count; the falsifiable alternative is that the plateau is flat in the decoding configuration, under which decorrelation would not be the lever the rule names. This requires model inference and is left to follow-on work; it is stated now so the decoding lever is testable rather than assumed.

5.4 A practitioner's reference

Everything needed to apply the lens fits on a page: a short checklist of what to do, a table of the formulas, and a reading of the common methods. The checklist comes first, and these are diagnostics, not universal defaults.

1. For evaluation, report n_{eff}, not just n: for the benchmark mean a budget of n per problem is worth $n/[1+(n-1)\rho_b]$ independent observations, at most $1/\rho_b$, so add problems rather than samples.
2. Estimate ρ_b from the log with Equation (15), clipped at zero on saturated benchmarks.
3. Distinguish the three goals: coverage keeps improving with n, selection plateaus at π_{mode} and can anti-scale, estimation saturates by $n \approx 1/\rho_b$.
4. For selection, lower the answer concentration (temperature, nucleus sampling, prompt diversity, model mixing) rather than raising n; sampling past the effective answer count buys nothing.
5. Separate the answer ceiling from the verifier ceiling: one is whether the correct answer is the mode, the other whether the right try is identifiable by the scorer.

A concrete pass through the checklist fixes the workflow. Take a benchmark on which the difficulty correlation is $\hat{\rho}_b \approx 0.5$ and the answers concentrate onto about a dozen modes per problem, the regime of the logs in Section 4.3. For evaluation, a budget of 256 samples per problem is worth only $n_{\text{eff}} \approx 1/\rho_b \approx 2$ independent observations of the benchmark mean, so a confidence interval computed as if the 256 were independent is too narrow by roughly elevenfold ($\sqrt{256/2}$); the honest move is to report n_{eff} and to spend fresh compute on fresh problems. For deployment with a verifier, the same 256 samples keep paying, since coverage carries no within-problem ceiling. For deployment without one, self-consistency has captured almost all it will within a few dozen samples, a small multiple of the effective answer count, and drawing the full 256 risks anti-scaling on the problems whose mode is wrong.

Three numbers drive the three goals: the difficulty correlation ρ_b, the effective answer count, and the verifier's accuracy. The first two are read off any sampling log already in hand by the estimators above, and the third is a property of the scorer. Reading them once states how much of a budget is real for each use, which is the whole of what the lens asks of a practitioner. Table 2 collects the formulas behind the checklist, each with the behavior it explains, from the design effect that opens the argument to the modal-hit rate that closes it.

Table 2: Formulas for correlated test-time scaling. Here n is the number of samples, s the per-attempt success probability, and ρ the intraclass correlation of the success indicators.

Quantity	Formula	Interpretation
Design effect	$d_{\text{eff}} = 1 + (n-1)\rho$	Correlation inflates the variance of the count.
Effective number of samples	$n_{\text{eff}} = n/[1 + (n-1)\rho]$	The usable count of independent draws.
Correlation ceiling	$n_{\text{eff}} \to 1/\rho$	Sampling cannot exceed $1/\rho$ effective draws.
Value of the n-th sample	$(1-\rho)/[1 + (n-1)\rho]^2$	Marginal worth decays as $1/(\rho n)^2$.
Coverage tax	$\text{pass@}n \leq 1 - (1-s)^n$	Correlation runs coverage below the independent curve.
Coverage tail	$\mathbb{P}[K{=}0] \sim \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\, n^{-\alpha}$	Heterogeneity gives power-law, not exponential, coverage.
Selection ceiling	$\pi_{\text{mode}} = \mathbb{P}_q[a_q^\star \text{ correct}]$	Self-consistency cannot beat the modal-hit rate.
Effective number of answers	$1/\sum_a p_q(a)^2$	Sets how fast plurality reaches its ceiling.
Majority-vote bound	$\mathbb{P}[\theta > \frac{1}{2}] \approx \Phi\big(\frac{s-1/2}{\sqrt{\rho s(1-s)}}\big)$	Lower bound on π_{mode}, loose when errors scatter.

The formulas price each behavior; reading the methods as estimands then assigns each its ceiling. Coverage escapes every ceiling because it asks only whether a correct sample exists; self-consistency and best-of-n meet the modal-answer wall because they read the answer distribution; and only benchmark-mean estimation meets the correlation ceiling, the one place the design effect of Section 3 truly binds.

Table 3: Test-time-scaling methods as estimands, with the ceiling that binds each.

Method	Reads	Estimand	Ceiling
Coverage / pass@n (with a verifier)	any correct sample	existence indicator	reachability, no $1/\rho$
Self-consistency / plurality	most common answer	mode of p_q	π_{mode}
Best-of-n (reward model)	top-scored sample	argmax of reward	reward-model ranking
Benchmark-mean estimate	success fraction $\hat{p}$	sample mean	$n_{\text{eff}} \leq 1/\rho_b$
Single sample / greedy	one answer	single draw	n/a

6 Conclusion

More sampling makes coverage climb while selection stalls; what separates them is the *identifiability gap*, the solvable problems whose answer a vote never returns. Selection stalls because it meets the modal ceiling π_{mode}, fixed by how often the most common answer happens to be right: on the dependent-draw log of Beeching et al. [20] the 256 tries per problem reduce to about thirteen, coverage reaches 0.88 while selection is right for only 0.45, and drawing more only sharpens a confident error. A different use meets a different limit: estimating a benchmark mean is capped by the correlation ceiling $1/\rho_b$, so with $\hat{\rho}_b \approx 0.4$–0.6 on

the released logs of Brown et al. [2], ten thousand attempts on a single problem buy the precision of about two, while the seed-to-seed term stays near zero ($\hat{\rho}_w \approx 0.0007$), leaving coverage with no within-problem ceiling. One sample count thus stands in for three different things at once: difficulty spread between problems, dependence between runs, and the collapse of answers onto a mode.

When to stop. Because each draw costs compute and both ceilings are low, the budget that pays is small and set by the goal: about $1/\rho_b$ samples to estimate a benchmark mean, on the order of the effective number of answers to select one, and no limit for coverage where a verifier can pick the correct sample out. Reporting the effective number of samples beside the nominal n, a closed form solved for rather than searched [32], says how much of a budget actually counts, and where it stops paying. The bottleneck in test-time scaling has moved from generating a correct answer to recognizing one: coverage shows the answers are already present, the modal ceiling shows that more sampling will not surface them, and the compute that extra draws cannot use is better spent correlating answers less and choosing among them better, not drawing more.

References

[1] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. 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.

[2] 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, 2024. arXiv:2407.21787.

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

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

[5] OpenAI. Openai o1 system card, 2024. arXiv:2412.16720.

[6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. arXiv:2501.12948.

[7] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. arXiv:2501.19393.

[8] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025. arXiv:2503.24235.

[9] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. arXiv:2107.03374.

[10] Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

[11] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977.

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

[13] 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.

[14] Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026. arXiv:2605.29800.

[15] Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight, 2025. arXiv:2502.04313.

[16] Natapong Nitarach. Model capability dominates: Inference-time optimization lessons from AIMO 3, 2026. arXiv:2603.27844.

[17] Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)?, 2025. arXiv:2502.17578.

[18] Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient prediction of pass@k scaling in large language models, 2025. arXiv:2510.05197.

[19] Noam Levi. A simple model of inference scaling laws, 2024. arXiv:2410.16377.

[20] Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models. Hugging Face blog, 2024. `https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute`.

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

[22] Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

[23] Yong Yi Bay and Kathleen A. Yearick. No 3D matrices: A unified tensor-product view of matrix-free Cartesian PDE solvers, 2026. arXiv:2606.25148.

[24] J. G. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society: Series B*, 10(2):257–261, 1948.

[25] Yong Yi Bay and Kathleen A. Yearick. Machine learning vs deep learning: The generalization problem. *arXiv preprint arXiv:2403.01621*, 2024. arXiv:2403.01621.

[26] Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? towards scaling laws of compound inference systems, 2024. arXiv:2403.02419.

[27] Tianchun Wang, Zichuan Liu, Yuanzhou Chen, Jonathan Light, Weiyang Liu, Haifeng Chen, Xiang Zhang, and Wei Cheng. On the effect of sampling diversity in scaling LLM inference, 2025. arXiv:2502.11027.

[28] 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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

[29] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.06452.

[30] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1904.09751.

[31] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1B LLM surpass 405B LLM? rethinking compute-optimal test-time scaling, 2025. arXiv:2502.06703.

[32] Yong Yi Bay and Kathleen A. Yearick. Solve for the hyperparameter, skip the search: Kolmogorov-optimal scaling laws for spline regression. *arXiv preprint arXiv:2606.23575*, 2026.

Appendix A: Elementary derivations

The body states the propositions and leans on these short calculations without pausing for them; each one underwrites a result above.

Beta-binomial moments and intraclass correlation. Take the latent rate $\theta \sim \text{Beta}(\alpha, \beta)$ with attempts $Y_i \mid \theta \sim$ i.i.d. Bernoulli(θ). The Beta has mean and variance

$$s = \mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}, \qquad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{s(1-s)}{\alpha+\beta+1}.$$

Two distinct attempts share only the rate θ, so their covariance is the variance of that shared rate: for $i \neq j$, $\text{Cov}(Y_i, Y_j) = \mathbb{E}[\theta^2] - s^2 = \text{Var}(\theta)$. Dividing by $s(1-s)$ gives the intraclass correlation of Equation (5),

$$\rho = \frac{\text{Var}(\theta)}{s(1-s)} = \frac{1}{\alpha+\beta+1},$$

which does not depend on n. Inverting, a target mean s and correlation ρ are realized by $\alpha = s(1-\rho)/\rho$ and $\beta = (1-s)(1-\rho)/\rho$, the map the figures use to set a Beta from (s, ρ).

Zero-success probability and its tail. The chance that none of n attempts succeeds is the Beta average of $(1-\theta)^n$, a standard beta integral:

$$\mathbb{P}[K=0] = \mathbb{E}_\theta\left[(1-\theta)^n\right] = \int_0^1 (1-\theta)^n \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\, d\theta = \frac{B(\alpha, \beta+n)}{B(\alpha,\beta)}.$$

Written with gamma functions this is $\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)} \cdot \frac{\Gamma(\beta+n)}{\Gamma(\alpha+\beta+n)}$. For large n the gamma ratio decays polynomially, $\Gamma(\beta+n)/\Gamma(\alpha+\beta+n) \sim n^{-\alpha}$ (Stirling, $\Gamma(x+a)/\Gamma(x+b) \sim x^{a-b}$), so

$$\mathbb{P}[K=0] \sim \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\, n^{-\alpha} \qquad (n \to \infty).$$

This is the power-law tail of Proposition 4: coverage approaches its limit polynomially, not exponentially.

The selection plateau. Plurality returns the most frequent answer, and the answer counts over n attempts are Multinomial(n, p_q). By the law of large numbers the empirical frequencies converge to p_q almost surely, so the empirical mode converges to the true mode $a_q^\star = \arg\max_a p_q(a)$, and per-problem plurality accuracy to $\mathbf{1}\{a_q^\star = c_q\}$. Averaging over problems gives the modal-hit rate π_{mode} of Equation (13).

The two-answer case recovers the majority vote. With only a correct answer and one error, write θ for the per-attempt correctness; then plurality is majority and $\hat{p} = K/n \to \theta$ almost surely, so in the limit the majority is right precisely if $\theta > \frac{1}{2}$. Its variance tends to $\text{Var}(\hat{p}) \to \rho\, s(1-s)$, and approximating θ as $\mathcal{N}\big(s, \rho\, s(1-s)\big)$ gives the closed form

$$\mathbb{P}[\theta > \tfrac{1}{2}] \approx \Phi\left(\frac{s - \frac{1}{2}}{\sqrt{\rho\, s(1-s)}}\right)$$

of Corollary 4. Because $\theta > \frac{1}{2}$ pushes every rival answer under $\frac{1}{2}$, hence under the correct one, this majority bound never exceeds π_{mode}.

Two-stage design effect. Pool attempts over problems, and let μ_i be the success rate of problem i, with between-problem correlation $\rho_b = \mathrm{Var}(\mu_i)/[s(1-s)]$ and a common within-problem correlation ρ_w given the problem. By the law of total covariance, two same-problem attempts $j \neq j'$ have

$$\mathrm{Cov}(Y_{ij}, Y_{ij'}) = \mathbb{E}\big[\rho_w \mu_i(1-\mu_i)\big] + \mathrm{Var}(\mu_i).$$

Since $\mathbb{E}[\mu_i(1-\mu_i)] = s(1-s) - \mathrm{Var}(\mu_i) = s(1-s)(1-\rho_b)$, the right-hand side is $s(1-s)\big[\rho_b + (1-\rho_b)\rho_w\big]$. Dividing by $s(1-s)$ gives the pooled correlation

$$\rho = \rho_b + (1-\rho_b)\rho_w$$

of Equation (14), and the same-problem design effect is $1 + (n-1)\rho$ [11].

Appendix B: Reproducibility

Every number, table, and figure in this work regenerates from a clean checkout. The propositions are checked numerically by `scripts/verify_math.py`, which verifies the identities of Sections 3–5 (including the variance identity (6), the coverage monotonicity of Proposition 2, the modal-answer selection ceiling (13) with its anti-scaling corollary and the majority-vote lower bound, and the two-stage decomposition (14)) against Monte Carlo simulation; all checks pass. The five model-based figures are generated by `scripts/make_figures.py` from closed forms and fixed-seed simulation; the correlated attempts use the de Finetti representation $\theta \sim \mathrm{Beta}(\alpha, \beta)$, $Y_i \mid \theta \sim \mathrm{Bernoulli}(\theta)$, with (α, β) set from (s, ρ) by Appendix A. The between-problem figure and Table 1 are produced by `scripts/analyze_brown.py`, which downloads the public sampling logs of Brown et al. [2], estimates $\hat{\rho}_b$ with a clustered (problem-level) bootstrap 95% interval (10^4 resamples, fixed seed), computes the coverage and self-consistency curves, and writes a small summary that the figure script reads. The within-problem figure and the $\hat{\rho}_w$, $\hat{\rho}_b$, and decomposition numbers of Section 4.3 are produced by `scripts/analyze_rhow.py`, which downloads the five-session best-of-n log of Beeching et al. [20], grades every completion with `math-verify` (the verifier behind the dataset's own labels, whose reported single-sample accuracy it reproduces to about a point), estimates the between- and within-problem correlations with problem-clustered bootstrap intervals, and computes the within-session coverage and plurality curves; the graded per-problem counts are cached so the figure rebuilds without re-downloading or re-grading the multi-hundred-megabyte log. The Python environment is pinned in `pyproject.toml` (`uv sync`); `make all` regenerates everything. Seeds are fixed, so the results are reproducible. The code, the cached result summaries, and the manuscript source are available at `https://github.com/bay-yearick-lab/sampling-ceilings`.