SUCCESS HAS A SHAPE: TailPass@k FOR REPEATED SAMPLING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated-sampling evaluation measures language models over multiple non-deterministic attempts on each task, but scalar summaries such as accuracy, pass@k, pass^k, or a single thresholded repeated-sampling score can conflate systems with different operational behavior. A model that occasionally finds a correct answer can have high discovery but low repeatability, while another model may have lower single-sample accuracy but much stronger consistency across attempts. We characterize failure modes of commonly used evaluation metrics and address information loss and uncertainty through a discovery–stability tail profile: for each task, reporting budget k, and threshold t, the profile is the probability that the model produces at least t correct responses among k attempts. We place a Dirichlet–categorical model on each task's latent outcome distribution, with binary exact match recovered as the Beta–Bernoulli special case, and propagate this uncertainty to dataset-level profiles (across tasks), yielding closed-form posterior means and covariances, together with posterior Monte Carlo credible intervals. The resulting uncertainty-aware report is TailPass@k. The framework recovers discovery, pass@k, stability, and single-sample accuracy as special cases and shows that accuracy is the area under the profile. Utilities over this profile combine its threshold probabilities into scalar scores that express preferences over discovery and repeatability, with uncertainty propagated to each score (e.g., power-moment utilities shift emphasis from discovery toward repeatability as their parameter λ increases). Exact enumeration of response patterns shows that endpoint metrics and uniform-threshold area can merge response patterns with the same total correctness but different discovery–stability structure, while the full profile exposes these distinctions.

1 INTRODUCTION

Test-time scaling has made repeated sampling a central part of language-model evaluation. A model may now be sampled many times for the same problem, searched over candidate reasoning traces, filtered by a verifier, aggregated by self-consistency, or evaluated as an agent across repeated interactions. Such protocols appear in code generation through pass@k (Chen et al., 2021), in reasoning through self-consistency and inference-scaling methods (Wang et al., 2023; Brown et al., 2024; Snell et al., 2025), in agent reliability benchmarks through pass^k (Yao et al., 2025), and in reasoning stability benchmarks through thresholded G-Pass@k_τ (Liu et al., 2025). In these settings, inference compute changes the event being measured.

A repeated-sampling benchmark produces a response bank, and scalar metrics project this bank onto different operational events. Average accuracy counts correct cells. pass@k asks whether at least one of k attempts succeeds, so it emphasizes discovery. Stability-style metrics ask whether many or all attempts succeed, so they emphasize repeatability. Majority and fixed-threshold scores lie between these endpoints. These summaries answer different operational questions. A model with rare success on many tasks can look strong under large-budget discovery while remaining weak under single-sample use; a model that solves fewer tasks reliably can show the opposite pattern. A scalar report can therefore reverse rankings as the budget changes, alias response banks with different success allocation, or hide uncertainty in close comparisons (Appendix K). These projection failures compound with measurement failures when the observed bit is verifier acceptance (Liu et al., 2023a;

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.

Stroebl et al., 2024). Dependent attempts and heterogeneous costs introduce further distinctions between the response-bank summary and deployment utility (Appendix K).

The relevant object is the *shape of the success-count* distribution. We introduce $\mathrm{TailPass@}k$ as a profile of discovery–stability for repeated-sampling evaluation. For a fixed sampling budget, each coordinate is the expected fraction of tasks reaching a required number of correct attempts, from one success through all k successes. The left edge recovers the discovery view associated with $\mathrm{pass@}k$; the right edge recovers stability; the interior measures how *rapidly* occasional success decays into reproducible success. A Beta–Bernoulli posterior quantifies uncertainty jointly across coordinates. Auxiliary responses, such as greedy decodes, can supply *prior evidence*, and a Dirichlet–categorical extension supports *categorical evaluation* by replacing binary exact-match outcomes with rubric labels for partial credit, formatting, efficiency, confidence, or other declared outcome scores. In agentic evaluation, these choices separate exploration-oriented systems that occasionally find solutions from exploitation-oriented systems that reproduce solutions reliably.

Across 760 task-specific model pairs, 138 reverse their finite-bank $\mathrm{pass@}k$ ranking as k varies, and large-budget discovery often disagrees with average accuracy. $\mathrm{TailPass@}k$ exposes discovery–repeatability trade-offs and repeatability gaps hidden by discovery scores. On AIME 2026, Qwen and GPT-OSS high differ by only 0.05 percentage points in discovery at $k = 16$, but by 19.74 points when all 16 attempts must succeed. On SuperGPQA, TailPass has the lowest mean threshold-selection regret among 11 rules, and TailPass-Moment has the highest mean held-out ranking agreement for $U_4(16)$, at all four tested fitting sizes (Appendix H). Posterior comparisons show that many point-estimate winners lack 95% posterior support. Greedy decoding provides useful small-sample prior evidence, while categorical rubrics induce large, interpretable rank changes under efficiency- and confidence-adjusted scoring. Convergence and Kendall's τ_b (Kendall, 1945) analyses quantify how quickly scalar views recover a full-bank reference ordering, while a temperature sweep shows that decoding temperature reshapes the discovery–stability spectrum.

We make four contributions:

1. failure-mode analyses of repeated-sampling metrics, covering ranking reversals, large-budget target mismatch, information loss from scalar summaries, and violations of measurement and sampling assumptions (Appendices K–M);
2. $\mathrm{TailPass@}k$, a discovery–stability profile with utility summaries that assign value to different correctness requirements (Sections 2–3);
3. joint posterior uncertainty (Section 3.1), changes in discovery and repeatability as the sampling budget grows (Section 2.2), discounted prior evidence from auxiliary responses (Appendix C.5), and categorical rubrics for partial credit, format, efficiency, and confidence (Section 3.3);
4. simulations of information loss and threshold-specific model selection with matched accuracy and endpoints (Appendix D), and mathematics and science experiments on profile shape, ranking uncertainty, prior transfer, rubric sensitivity, convergence, rank correlation, and decoding temperature (Section 4; Appendices G.1–G.4, H, and I.1).

2 Shape of Success

Repeated-sampling metrics summarize different events in a response bank. Discovery–stability profiles retain the threshold probabilities from which discovery, repeatability, and utility scores are derived.

2.1 Response-bank Shape

Let $R \in \{0,1\}^{M\times N}$ be the binary response bank for M questions and N attempts per question, with $R_{ij} = 1$ for a correct response. The row counts $\gamma_i = \sum_j R_{ij}$ and histogram $h_c = \#\{i : \gamma_i = c\}$, $c = 0, \ldots, N$, describe how successes are distributed across questions. Accuracy, $(MN)^{-1}\sum_i \gamma_i$, retains only their mean: successes may be spread across many questions or concentrated on a few. Under the exchangeable-attempt abstraction, the histogram is the finite-bank shape; answer identities and attempt order require the underlying responses. At budget $k = N$, its survival curve is the

finite-bank threshold profile (Appendix C.1). Even banks with equal accuracy and both endpoint scores can differ at intermediate thresholds (Appendix D.1).

Those differences matter for retry and reliability: equal scores at one operating point can conceal different probabilities of finding a success at another budget or meeting a repeated-success threshold. Same-accuracy comparisons isolate the allocation of successes across questions, since the total correctness is held fixed.

2.2 The Discovery–Stability Tail Surface

A finite bank provides evidence about a prospective sampling process. Let p_i be the probability that an independent attempt on question i is correct. For a fresh budget of k attempts, $X_{i,k} \mid p_i \sim \mathrm{Binomial}(k, p_i)$, and the tail kernel is

$$q_{k,t}(p) = \Pr\{\mathrm{Binomial}(k,p) \geq t\} = \sum_{r=t}^{k} \binom{k}{r} p^r (1-p)^{k-r}, \qquad t = 1, \ldots, k. \tag{1}$$

The question-level coordinate is $S_{i,k,t} = q_{k,t}(p_i)$; its dataset average

$$S_{k,t} = \frac{1}{M} \sum_{i=1}^{M} q_{k,t}(p_i) \tag{2}$$

is the expected fraction of questions reaching threshold t. The binary $\mathrm{TailPass@}k$ profile is $S_k = (S_{k,1}, \ldots, S_{k,k})$, with question-level profile $\mathrm{TailPass}_i\mathrm{@}k := S_{i,k} = (S_{i,k,1}, \ldots, S_{i,k,k})$. Across budgets, these profiles form a tail surface indexed by compute budget k and required number of correct attempts t. The observed-run count N determines the available evidence and need not equal the reporting budget k.

The left endpoint $S_{k,1}$ is the population analogue of $\mathrm{pass@}k$; the right endpoint $S_{k,k}$ is the probability of all k attempts succeeding, averaged over questions. Intermediate thresholds measure partial repeatability and include thresholded $\mathrm{G\text{-}Pass@}k_\tau$. The derivative of $q_{k,t}$ is a $\mathrm{Beta}(t, k-t+1)$ density. For $k > 1$, low, middle, and high thresholds emphasize different regions of the success-probability axis (Appendix C.2).

Operational features. The normalized profile $S_k(\rho) = S_{k,\lceil k\rho \rceil}$, $\rho \in (0, 1]$, approaches the survival curve of the latent p_i as k grows, at continuity points of their empirical distribution (Appendix C.6). Its area is average single-sample accuracy, $k^{-1} \sum_{t=1}^{k} S_{k,t} = M^{-1} \sum_i p_i$, by the tail-sum identity (Billingsley, 1995). Thus equal-area profiles can differ in discovery, stability, and the decay between them. The discovery endpoint approaches coverage, the fraction of questions with $p_i > 0$, rather than their average success probability. A question with rare success can therefore contribute nearly one to large-budget discovery while contributing little to accuracy or high-threshold repeatability.

2.3 Failure Modes as Missing Shape

Finite-bank $\mathrm{pass@}k$ retains one budget-dependent coordinate: early budgets emphasize average success mass, large budgets emphasize observed coverage, and intermediate budgets mix higher-order row-count moments. Rankings can therefore reverse, and distinct banks can remain indistinguishable (Appendix Sections K.1–K.3). Likewise, $\mathrm{G\text{-}Pass@}k_\tau$, strict majority, and pass^k select coordinates, while $\mathrm{mG\text{-}Pass@}k$ and $\mathrm{AUC@}K$ average over threshold or budget axes; modal-answer majority additionally requires answer clusters and tie-breaking (Appendix M). Retaining the profile exposes these projection losses, but does not correct unreliable verification, dependent attempts, or unmodeled failure costs (Appendix K).

3 $\mathrm{TailPass@}k$: Uncertainty-aware Discovery–Stability Evaluation

An uncertainty-aware $\mathrm{TailPass@}k$ report consists of posterior summaries of S_k and pre-specified scalar functionals of the same profile. The latent profile describes fresh independent attempts; its observed-bank analogue draws without replacement and recovers finite-sample $\mathrm{pass@}k$ at $t = 1$

(Appendix C.1). The latent probability p_i differs from the observed fraction γ_i/N. For $N \geq k$, the hypergeometric finite-bank profile is an unbiased estimator of the latent profile under independent Bernoulli sampling, but its realized value depends on the observed responses. When $k > N$, sampling without replacement is undefined; a prospective profile instead requires the sampling model and corresponding assumption and prior checks.

3.1 Bayesian Uncertainty over the Profile

Independent priors $p_i \sim \mathrm{Beta}(\alpha_i^0, \beta_i^0)$ propagate uncertainty through the tail map. Optional prior evidence is a separate response bank $R^0 \in \{0,1\}^{M \times D}$ with counts $\gamma_i^0 = \sum_d R_{i,d}^0$, following prior-count augmentation in Bayesian LLM evaluation (Hariri et al., 2026c). With transfer weight $\eta \in [0,1]$, set $\alpha_i^0 = 1 + \eta\gamma_i^0$ and $\beta_i^0 = 1 + \eta(D - \gamma_i^0)$. Without auxiliary evidence, $D = 0$ gives the uniform $\mathrm{Beta}(1,1)$ prior. Bernoulli conjugacy yields

$$p_i \mid R_i, R_i^0, \eta \overset{\text{ind}}{\sim} \mathrm{Beta}(a_i, b_i), \qquad a_i = \alpha_i^0 + \gamma_i, \quad b_i = \beta_i^0 + N - \gamma_i. \tag{3}$$

The posterior mean profile averages beta-binomial predictive tails:

$$\mu_{k,t} := \mathbb{E}[S_{k,t} \mid R, R^0, \eta] = \frac{1}{M} \sum_{i=1}^{M} \sum_{r=t}^{k} \binom{k}{r} \frac{\mathrm{B}(a_i + r, b_i + k - r)}{\mathrm{B}(a_i, b_i)}, \qquad t = 1, \ldots, k, \tag{4}$$

where B is the Beta function. Joint uncertainty follows by drawing $p_i^{(h)} \sim \mathrm{Beta}(a_i, b_i)$ and evaluating every threshold and scalar summary on the same draw $p_{1:M}^{(h)}$. Componentwise equal-tailed credible intervals (Gelman et al., 2013), closed-form covariance, and the evaluation algorithm appear in Appendices C.4 and C.3 and Algorithm 1.

These intervals quantify uncertainty in the latent expected fraction of questions meeting a threshold, conditional on the observed questions and the prior. A realized future benchmark adds sampling variation even if every p_i is known; predictive intervals for that benchmark require the additional future-sampling layer (Appendix C.1). Shared posterior draws preserve dependence among profile coordinates and utilities. The reported coordinatewise intervals have the stated posterior mass for each coordinate separately; simultaneous coverage of the whole profile requires a joint credible-set construction.

Auxiliary evidence affects every threshold through the same posterior on p_i. The case $\eta = 1$ retains its full counts, whereas $\eta = 0$ excludes them. The prior rule is fixed across utility views; changing a threshold preference does not supply new evidence about success probabilities. Appendix C.5 gives prior sensitivity and calibration procedures based on predicting held-out success counts.

3.2 Expected-Utility Scalarization: TailPass@k

A linear utility assigns marginal value $\omega_{t,k}$ to the t-th correct sample:

$$U_\omega(k) = \sum_{t=1}^{k} \omega_{t,k} S_{k,t}, \qquad \omega_{t,k} \geq 0, \quad \sum_{t=1}^{k} \omega_{t,k} = 1.$$

This weighted-sum scalarization (Marler & Arora, 2004) commutes with the average over questions. Endpoint weights recover discovery and stability; uniform weights recover single-sample accuracy.

Equivalently, the weights define a nondecreasing payoff $u_k(x) = \sum_{t=1}^{x} \omega_{t,k}$ for x correct attempts. Requiring at least r successes assigns all weight to $S_{k,r}$; equal value per correct attempt gives uniform weights. Thus the utility depends on the operational consequence of success, not on which model ranks first. Realizing discovery value also requires identifying a correct candidate: a deployed verifier or voting policy needs its own evaluation (Appendix L.1).

For $\lambda > 0$, the power-moment weights $\omega_{t,k}^{(\lambda)} = (t/k)^\lambda - ((t-1)/k)^\lambda$ give

$$U_\lambda(k) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}\left[(X_{i,k}/k)^\lambda \mid p_i\right].$$

Small λ emphasizes discovery thresholds, $\lambda = 1$ gives accuracy, and larger λ shifts value toward high-repeatability thresholds. The exponent λ and budget k encode outcome preferences and operating constraints and are fixed independently of test rankings (Appendix E). Appendix C.8 gives posterior moments, alternative weight families, and nonlinear summaries, including the question-first quadratic RMS spectrum score (QRS). Unlike linear utilities, nonlinear summaries can depend on whether questions are averaged before or after scalarization.

3.3 Rubric-aware Categorical Evaluation

Let $C_{i,j} \in \{1, \dots, L\}$ be an attempt's label under a pre-specified rubric, with latent label probabilities π_i and scores $v = (v_1, \dots, v_L) \in [0,1]^L$. A Dirichlet prior is conjugate to the categorical observations. For future counts $Y_{i,k} \mid \pi_i \sim \mathrm{Multinomial}(k, \pi_i)$, define the normalized bank score $A^v_{i,k} = k^{-1} v^\top Y_{i,k}$ and tail profile

$$T^v_{i,k}(\rho) = \Pr\{A^v_{i,k} \geq \rho \mid \pi_i\}, \qquad T^v_k(\rho) = \frac{1}{M} \sum_i T^v_{i,k}(\rho), \quad \rho \in [0,1].$$

Wrong/correct labels with $v = (0,1)$ recover the binary profile at $\rho = t/k$ and the Beta–Bernoulli model with $p_i = \pi_{i,2}$. Partial credit, format, confidence, and efficiency rubrics differ in their category map and score vector, both fixed before ranking models. The category map specifies which distinctions the evaluation retains, while v_c assigns credit to label c. Changing either is a rubric sensitivity analysis over declared evaluation choices. The binary construction therefore extends to finite rubric schemata without changing the interpretation of a profile coordinate as the probability of meeting an outcome requirement.

The posterior mean is a Dirichlet–multinomial predictive tail; shared Dirichlet draws propagate uncertainty across thresholds and summaries. The categorical experiments use moment utilities $U_{v,\lambda}(k) = M^{-1} \sum_i \mathbb{E}[(A^v_{i,k})^\lambda \mid \pi_i]$: small λ emphasizes positive rubric credit, $\lambda = 1$ gives the expected per-attempt score, and larger λ emphasizes high scores. Appendix C.7 gives their threshold-integral form, predictive and covariance formulas, and Algorithm 2.

4 Experiments

We evaluate $\mathrm{TailPass@}k$ against scalar repeated-sampling metrics using competition-mathematics response banks. The primary banks contain 20 model configurations on AIME'24, AIME'25, HMMT'25, and BrUMO'25, with 30 questions per benchmark and 80 stochastic attempts per question. Sampling uses temperature 0.6 and top-$p = 0.95$. The aggregate analysis weights the 120 questions equally. Binary analyses retain all 20 configurations; categorical analyses retain 19 after excluding one with nonfinite prompt log-probabilities. The comparisons vary sampling budgets and correctness thresholds, quantify posterior ranking uncertainty, and assess greedy prior evidence and categorical rubrics (Appendix J). Additional Math and SuperGPQA banks (M-A-P Team et al., 2025) extend the profile comparisons across mathematics and science. These banks contain 186 mathematics questions and 3,600 SuperGPQA questions, with 50 questions in each of 72 fields. They evaluate four configurations: one Qwen3.6-35B-A3B configuration and three GPT-OSS-20B reasoning-effort settings (low, medium, high), using 80 attempts per configuration and question. Their decoding settings differ from the primary banks; comparisons retain the source correctness labels and concern the recorded configurations. On SuperGPQA, disjoint fitting and evaluation attempts provide held-out tests of threshold-specific model selection and utility-ranking agreement (Appendix H).

4.1 Response-bank Diagnostics

Across the $4\binom{20}{2} = 760$ task-specific model pairs in the response-bank diagnostics (Appendix J.5), 138 pairs reverse their finite-bank $\mathrm{pass@}k$ ranking over $k = 1{:}80$. Since each bank has $N = 80$ samples, $\mathrm{pass@}80$ equals observed coverage, and its strict ordering disagrees with average accuracy in 83 of 642 non-tied comparisons. At $k = 16$, $\mathrm{pass@}16$ disagrees with $\mathrm{G\text{-}Pass@}16_{0.5}$ on 98 pairs, $\mathrm{Maj@}16$ on 107 pairs, pass^{16} on 157 pairs, $\mathrm{mG\text{-}Pass@}16$ on 129 pairs, and $\mathrm{AUC@}16$ on 30 pairs. These metrics project different parts of the threshold profile analyzed in Section 2.2 and Appendix M.

An HMMT'25 pair illustrates the reversal. gpt-oss-medium has higher discovery, $S_{16,1} = 0.822$ versus 0.727, and the posterior probability that Qwen3-Thinking is larger at this endpoint is below 0.001.

Table 1: Posterior profile endpoints $S_{16,t}$ and moment utilities $U_\lambda(16)$ for a discovery–stability reversal on HMMT'25. Values are posterior means with 95% credible intervals. Model short names are expanded in Table 35.

Score	Operating view	Qwen3-Thinking	gpt-oss-medium	Pr(Qwen3-Thinking > gpt-oss-medium)
$S_{16,1}$	Discovery endpoint	0.727 [0.695, 0.760]	0.822 [0.793, 0.853]	< 0.001
$U_{\lambda=.25}(16)$	Discovery-weighted	0.626 [0.607, 0.646]	0.665 [0.646, 0.685]	0.003
$U_{\lambda=1}(16)$	Uniform utility	0.493 [0.483, 0.503]	0.437 [0.423, 0.450]	> 0.999
$U_{\lambda=4}(16)$	Stability-weighted	0.382 [0.366, 0.396]	0.239 [0.224, 0.255]	> 0.999
$S_{16,16}$	Strict stability endpoint	0.273 [0.238, 0.303]	0.128 [0.104, 0.149]	> 0.999

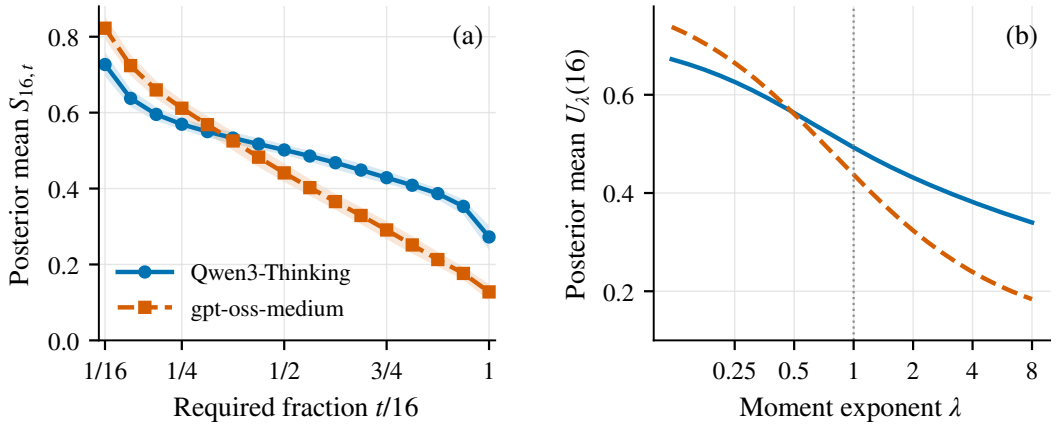

Figure 1: Discovery–stability comparison on HMMT'25 at $k = 16$. (a) Posterior means of the TailPass@16 coordinates $S_{16,t}$, with pointwise 95% credible bands. gpt-oss-medium has the larger discovery endpoint; Qwen3-Thinking has the larger moderate- and high-threshold tails. (b) Posterior mean utility $U_\lambda(16)$ as the moment exponent varies; the dotted line marks $\lambda = 1$. Discovery-weighted utility at $\lambda = 0.25$ favors gpt-oss-medium, while uniform and stability-weighted utilities at $\lambda = 1$ and 4 favor Qwen3-Thinking.

With repeatability weighted more heavily, the ordering flips: Qwen3-Thinking has higher $U_{\lambda=1}(16)$, higher $U_{\lambda=4}(16)$, and higher strict stability $S_{16,16}$, each with posterior probability above 0.999 for the Qwen3-Thinking ordering (Table 1 and Figure 1). The discovery-heavy utility $U_{\lambda=.25}(16)$ favors gpt-oss-medium.

In the controlled simulations, TailPass recovers model choices that depend on the intermediate threshold even when accuracy and both endpoints match exactly. Its posterior-mean profiles reduce regret relative to rankings fixed across thresholds (Appendix D.2).

Within-task rankings show the same dependence on the declared utility. pass@16 disagrees with $U_{\lambda=.25}(16)$ on 42 of 760 task-specific model pairs, with $U_{\lambda=1}(16)$ on 96 of 759 non-tied pairs, and with $U_{\lambda=4}(16)$ on 137 of 760 pairs. Changing the moment parameter also changes strict rankings within the moment-utility family: $U_{\lambda=.25}(16)$ and $U_{\lambda=4}(16)$ disagree on 97 pairs, $U_{\lambda=.25}(16)$ and $U_{\lambda=1}(16)$ disagree on 54 pairs, and $U_{\lambda=1}(16)$ and $U_{\lambda=4}(16)$ disagree on 46 pairs. Posterior uncertainty changes which strict winners are separable. Under the TailPass posterior at $k = 16$, 151 model pairs have less than 95% posterior support for the larger posterior-mean discovery score $S_{16,1}$. The corresponding counts are 87 for $U_{\lambda=.25}(16)$, 51 for $U_{\lambda=1}(16)$, and 69 for $U_{\lambda=4}(16)$. Strict point estimates therefore leave some pairwise decisions unsupported. Appendix I gives the required response information, pairwise diagnostic counts, and additional profile examples.

4.2 Uncertainty in TailPass@k & Utility Rankings

Under the TailPass@8 uncertainty protocol in Appendix J.4, interval-aware rankings use the moment-family utility $U_{\lambda=2}(8)$ and group adjacent models when the 95% credible interval for their pairwise scalar difference contains zero. Rank agreement is measured with Kendall's τ_b. Strict posterior-mean rankings are more granular than the posterior evidence. On the aggregate bank, 8 of the 190

Table 2: Interval-aware rankings by utility $U_{\lambda=2}(8)$. The strict reference ranking for each task is the all-80-trial posterior-mean ranking under the same utility. A pair is unresolved when the 95% equal-tailed posterior credible interval for the pairwise utility difference contains zero. Rank groups are the transitive closure of unresolved adjacent comparisons after sorting by posterior mean. A UQ block is one such uncertainty-aware rank group; the largest UQ block column reports its strict-rank span. Kendall's τ_b compares the grouped ranking with the strict reference ranking.

Task	Unresolved pairs	Rank groups	Largest group	τ_b	Largest UQ block
All	8/190	14	4	0.973	strict ranks 7–10
AIME'24	14/190	11	5	0.951	strict ranks 6–10
AIME'25	21/190	8	7	0.915	strict ranks 7–13
HMMT'25	13/190	11	5	0.957	strict ranks 8–12
BrUMO'25	16/190	10	6	0.938	strict ranks 3–8

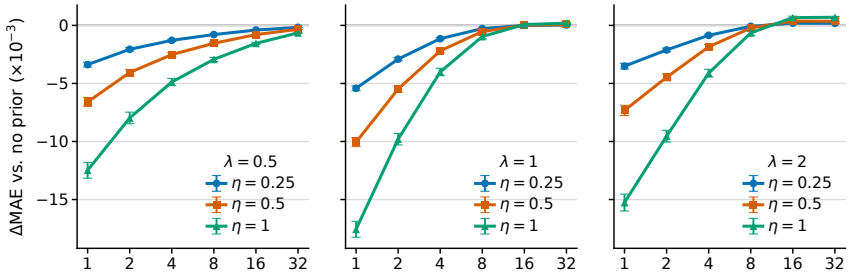

Figure 2: Effect of greedy prior transfer on held-out prediction error for aggregate U_λ@16 summaries. Values are paired changes in mean absolute error relative to the no-prior posterior across the number of observed top-p trials per question; negative values indicate that the greedy prior improves prediction of the held-out top-p finite-bank reference. Error bars are ± 1.96 standard errors over the matched model–subsample units.

model pairs are unresolved under $U_{\lambda=2}(8)$, and strict ranks 7 through 10 form one uncertainty-aware rank group. The task-specific rankings vary in severity: AIME'25 has the most unresolved all-pair comparisons (21/190) and eight rank groups, while BrUMO'25 has fewer unresolved comparisons (16/190) but still compresses strict ranks 3 through 8 into a shared block.

Profile uncertainty is not uniform over thresholds. The aggregate bank is most certain near the middle thresholds and widest at the stability endpoint (0.024 at $t = 8$ versus 0.016 at $t = 4$). AIME'24 and BrUMO'25 show stronger stability-endpoint widening, with median widths 0.057 and 0.054 at $t = 8$. HMMT'25 has its widest coordinate at discovery (0.051 at $t = 1$). Scalar uncertainty changes with the threshold weights used to summarize the same posterior profile. On the aggregate bank, the median 95% interval width for $U_\lambda(8)$ is 0.012 at $\lambda = 1$, 0.017 at $\lambda = 4$, and 0.020 at $\lambda = 8$. The unresolved-pair count is also viewpoint-dependent: 7/190 pairs are unresolved at $\lambda = 1$, 11/190 at $\lambda = 4$, and 13/190 at $\lambda = 8$. The leading aggregate model is separated from the runner-up at every tested λ. Appendix G.2 gives the full rankings, figures, profile-width table, and scalar-view sensitivity table; Appendix J.4 gives the reproducibility protocol. Appendix L.4 analyzes the analogous finite-bank false-ranking mechanism for pass@k.

4.3 Prior Evidence from Greedy Decoding

The prior-count construction in Section 3.1 allows a deterministic auxiliary run to enter the Beta–Bernoulli posterior as discounted evidence. We test this mechanism by using a single greedy decode as a one-trial prior for the top-p stochastic response bank. Transfer weights $\eta \in \{0, .25, .5, 1\}$ range from no transfer to full one-trial transfer, and the validation predicts a held-out top-p finite-bank reference from $n \in \{1, 2, 4, 8, 16, 32\}$ observed stochastic trials per question. The scalar summaries are U_λ@16 with $\lambda \in \{.5, 1, 2\}$, covering discovery-weighted, uniform, and repeatability-weighted views. The full protocol is in Appendix J.3.

Table 3: Greedy-prior sensitivity for moment utilities $U_\lambda(16)$ on the aggregate response bank. Predictions use n_{obs} observed top-p trials per question and a held-out finite-bank reference at $k = 16$; intervals are clipped normal 95% posterior intervals. Negative ΔMAE indicates improvement over the no-prior baseline.

		$n_{\text{obs}} = 4$			$n_{\text{obs}} = 16$		
λ	η	MAE	ΔMAE	95% cover / width	MAE	ΔMAE	95% cover / width
0.5	0	0.0590	0.0000	0.323/0.0526	0.0299	0.0000	0.309/0.0302
0.5	1	0.0541	−0.0049	0.325/0.0490	0.0283	−0.0015	0.315/0.0295
1	0	0.0429	0.0000	0.342/0.0552	0.0152	0.0000	0.525/0.0276
1	1	0.0389	−0.0040	0.362/0.0502	0.0153	0.0001	0.490/0.0268
2	0	0.0691	0.0000	0.195/0.0603	0.0266	0.0000	0.304/0.0319
2	1	0.0650	−0.0041	0.228/0.0551	0.0273	0.0007	0.301/0.0310

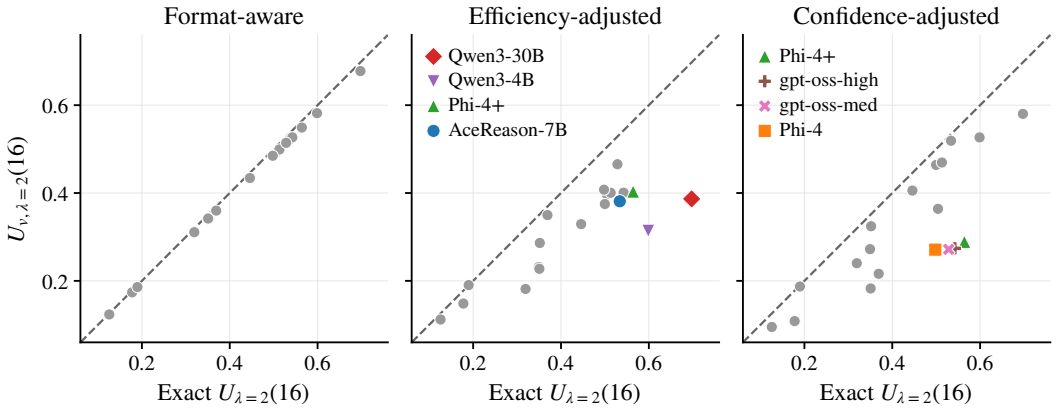

Figure 3: Aggregate rubric utilities $U_{v,\lambda=2}(16)$ against binary exact-correctness utility $U_{\lambda=2}(16)$. Horizontal coordinates are posterior means under exact correctness, and vertical coordinates are posterior means under the indicated rubric v. Legends identify the highlighted models; the remaining models are gray. Points below the diagonal have lower posterior means under the categorical analysis; efficiency and confidence produce the largest deviations.

Figure 2 and Table 3 show that greedy transfer helps most when only a few stochastic samples are available. With $n = 4$ observed top-p trials, full transfer ($\eta = 1$) reduces aggregate held-out MAE from 0.0590 to 0.0541 for $\lambda = .5$, from 0.0429 to 0.0389 for $\lambda = 1$, and from 0.0691 to 0.0650 for $\lambda = 2$. At $n = 16$, the discovery-weighted summary still improves ($\Delta\text{MAE} = -0.0015$ for $\lambda = .5$), but the uniform and repeatability-weighted summaries slightly worsen ($\Delta\text{MAE} = 0.0001$ and 0.0007). Greedy decoding supplies useful low-sample evidence for all three utilities, with the improvement persisting to sixteen observed trials for the discovery-weighted summary. Appendix G.3 gives the full η grid, per-task results, interval diagnostics, prior-alignment checks, and full-bank sensitivity; Appendix G.3.2 shows the same sign pattern for $k \in \{2, 4, 8, 16\}$.

4.4 Rubric-aware Categorical Evaluation

Under the categorical protocol in Appendix J.6, changing the rubric and its category prior changes the preferred systems (Appendix Table 21). Exact and format-aware utilities share the same aggregate top five. Efficiency-adjusted utility ranks gpt-oss-medium first at 0.466 [0.459, 0.472] and Phi-4 second, while Qwen3-Thinking and Qwen3-4B fall to seventh and twelfth. Confidence-adjusted utility keeps Qwen3-Thinking and Qwen3-4B first and second but moves AR-Nemotron to third. The induced rank changes reflect the declared operating criterion: concise or lower-perplexity correct attempts receive more credit than verbose or higher-perplexity correct attempts.

In Figure 3, format-aware points lie close to the diagonal. Efficiency-adjusted scores discount long correct completions, especially for the Qwen3 models, while concise models such as gpt-oss-medium and Phi-4 move upward in rank. Confidence-adjusted scores separate models with similar exact-

match utilities by posterior mass on lower-perplexity correct attempts. Appendix Figures 16 and 17 show that these qualitative changes persist across test-time budgets and moment-family sensitivity settings.

5 Related Work

Repeated sampling is now a standard way to spend inference compute in language-model evaluation. Code-generation benchmarks use $\mathrm{pass@}k$ for any-correct success among k sampled programs (Chen et al., 2021; Li et al., 2022; Liu et al., 2023a), with a combinatorial finite-sample estimator introduced by Chen et al. (2021). Reasoning methods aggregate sampled answers through self-consistency (Wang et al., 2023), select candidates with learned verifiers (Cobbe et al., 2021; Lightman et al., 2024), search over reasoning states (Yao et al., 2023), and scale inference computation (Brown et al., 2024; Snell et al., 2025). Surveys organize these inference regimes and their evaluation (Zhang et al., 2025; Hariri et al., 2026a). Agent reliability benchmarks report all-success scores through τ-bench (Yao et al., 2025), while $\mathrm{G\text{-}Pass@}k_\tau$ measures reasoning stability at a specified correctness threshold (Liu et al., 2025). These protocols expose response banks but often reduce them to one-dimensional any-success, majority, selected-candidate, or average scores. Other protocols change the success label: reasoning-faithfulness criteria may require both a correct chain of thought and a correct final answer (Wen et al., 2025b), while repeated safety evaluations estimate failure probabilities under sustained inference (Broadwater, 2026). ReasonBENCH measures variation in reasoning quality and cost across repeated executions (Potamitis et al., 2025). Repeated-sampling scores also require statistical care. Bayesian evaluation methods quantify uncertainty in binary LLM outcomes (Longjohn et al., 2025) and in rubric-weighted scores and model rankings (Hariri et al., 2026c;b); TailPass places the posterior on thresholded success events, so scalar rankings depend on declared threshold weights. Rubric-aware evaluation and LLM-as-judge methods provide labels for open-ended outputs (Zheng et al., 2023; Liu et al., 2023b), while verifier, selector, and rubric error remain separate measurement problems. Appendix F situates these distinctions among adjacent repeated-sampling and utility-based evaluation schemes.

6 Conclusion: Strengths, Limitations & Future Directions

$\mathrm{TailPass@}k$ is a shape-aware profile for repeated-sampling evaluation. It reports the *success shape* as a full tail profile, covering the spectrum from one discovered success to reproducible success across all k attempts. The uncertainty-aware report attaches posterior uncertainty to every coordinate; auxiliary measurements (e.g., greedy decoding) provide *prior* evidence; and the Dirichlet–categorical extension supports *rubric-aware* utilities. Posterior inference assumes independent attempts and the declared labels; its uncertainty does not correct verifier errors or establish whether a utility reflects deployment needs.

The experiments show that scalar rankings depend on the declared utility as well as the response bank. Moment-family utilities provide a compact sensitivity path: small λ emphasizes low discovery thresholds, $\lambda = 1$ gives the uniform threshold area, and larger λ emphasizes repeatability. This choice is empirically consequential. Response-bank diagnostics show reversals between $\mathrm{pass@}16$ and moment utilities; uncertainty-aware rankings show that unresolved pairwise comparisons change with λ; prior-transfer validation shows that greedy evidence is most useful for low-sample, discovery-weighted summaries; and categorical experiments show that changing the category map, rubric scores, and moment parameter can alter model orderings (Sections 4.1, 4.2, 4.3, and 4.4).

In the controlled simulations, TailPass posterior-mean profiles reduce model-selection regret relative to fixed scalar rankings even when accuracy and both endpoints agree (Appendix D). The profile should therefore remain the primary reported object, with scalar utilities treated as pre-specified operating views of the same posterior profile. The moment family expresses the discovery–repeatability trade-off through one interpretable parameter. Future work can develop and validate utility families for deployment-specific costs, safety labels, latency, verifier and selector outcomes, adaptive sample allocation, and partial-credit rubrics (Appendices G.3, G.2, and G.4).

A complementary direction is to study how sampling protocols reshape $\mathrm{TailPass@}k$: the temperature sweep in Appendix I.1 shows that decoding reshapes the discovery–stability profile, and similar analyses can compare other sampling strategies and adaptive test-time scaling.

AI USE STATEMENT

We used commercial large language models and a local in-house model to assist with writing code and code documentation, preparing figures, polishing the manuscript, summarizing research content, and identifying related work. These tools also supported research ideation, checking the authors' mathematical proofs and theoretical analysis, and critical review of the study's contributions. The authors take full responsibility for the final text, theoretical claims, experimental results, and research artifacts, including all AI-assisted contributions.

REPRODUCIBILITY STATEMENT

The 20-configuration response banks use 80 trials per question, sampled at temperature 0.6 with top-$p = 0.95$ and seeds 1234 through 1313 (Appendix J.2). Model identifiers, benchmark sources, prompt templates, and inference settings are specified in Appendix J, together with experiment-specific priors, evaluation budgets, resampling procedures, posterior draw counts, and interval definitions. The Math and SuperGPQA evaluations retain source correctness labels and aligned question identities (Appendix H.1). SuperGPQA model comparisons use disjoint fitting and held-out attempts, with the same partitions reused across methods (Appendix H.3).

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report. arXiv preprint arXiv:2504.21318, 2025. URL https://arxiv.org/abs/2504.21318.

Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, Joonwon Jang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Euisoon Kim, Hyosang Kim, Jihoon Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Gwangho Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Young Min Paik, Yongmin Park, Youngyong Park, Sanghyun Seo, Sihoon Yang, Heuiyeen Yeen, Sihyuk Yi, and Hyeongu Yun. EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes. arXiv preprint arXiv:2507.11407, 2025. URL https://arxiv.org/abs/2507.11407.

Bespoke Labs. Bespoke-Stratos: The unreasonable effectiveness of reasoning distillation. Blog post, 2025. URL https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation. Accessed 2026-09-24.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, third edition, 1995. ISBN 978-0-471-00710-4. URL https://books.google.com/books/about/Probability_and_Measure.html?id=z39jQgAACAAJ.

Keita Broadwater. Evaluating LLM safety under repeated inference via accelerated prompt stress testing. arXiv preprint arXiv:2602.11786, 2026. URL https://arxiv.org/abs/2602.11786.

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. URL https://arxiv.org/abs/2407.21787.

Brown University Math Olympiad Organizers. Brown University Math Olympiad (BrUMO) 2025 archive. Official BrUMO archive, 2025. URL https://www.brumo.org/archive. Accessed 2026-09-24.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374. Version 2.

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

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

FuseAI. FuseO1-DeepSeekR1-QwQ-SkyT1-Flash-32B-Preview. Hugging Face model card, 2025. URL https://huggingface.co/FuseAI/FuseO1-DeepSeekR1-QwQ-SkyT1-Flash-32B-Preview. Accessed 2026-09-24.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013. ISBN 978-1-4398-4095-5. doi: 10.1201/b16018.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Tyler Griggs, Shiyi Cao, Dacheng Li, Shu Liu, Shishir G. Patil, Matei Zaharia, Joey Gonzalez, and Ion Stoica. Think less, achieve more: Cut reasoning costs by 50% without sacrificing accuracy. Project blog post, 2025. URL https://novasky-ai.github.io/posts/reduce-overthinking/. Accessed 2026-09-24.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. OpenThoughts: Data recipes for reasoning models. arXiv preprint arXiv:2506.04178, 2025. URL https://arxiv.org/abs/2506.04178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, Cambridge, second edition, 1952. ISBN 978-0-521-35880-4. URL https://books.google.com/books/about/Inequalities.html?id=t1RCSP8YKt8C. First paperback edition, 1988.

Mohsen Hariri. Scorio lite. Hugging Face dataset releases, 2026. URL https://huggingface.co/datasets/harimo/scorio-lite. Scorio collection: https://huggingface.co/datasets/harimo/scorio. Math and GPQA companions contain the same attempts as Scorio Lite.

Mohsen Hariri, Weicong Chen, Nahal Shahini, Vikash Singh, Kai Ye, Amirhossein Samandar, Debargha Ganguly, Sreehari Sankar, Yanyan Zhang, Shouren Wang, Jerry Peng, Biyao Zhang, Michael Hinczewski, and Vipin Chaudhary. Test-time scaling in reasoning LLMs: Inference regimes, evaluation, and reproducibility. arXiv preprint arXiv:2608.04001, 2026a. URL https://arxiv.org/abs/2608.04001.

Mohsen Hariri, Michael Hinczewski, Jing Ma, and Vipin Chaudhary. Ranking reasoning LLMs under test-time scaling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33437–33478, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1544. URL https://aclanthology.org/2026.acl-long.1544/.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't Pass@k: A Bayesian framework for large language model evaluation. In *International Conference on Learning Representations*, 2026c. doi: 10.48550/arXiv.2510.04265. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f04edfc65463d020629673a4bc4c58e7-Abstract-Conference.html.

Harvard–MIT Mathematics Tournament. HMMT february 2025 archive: Problems and solutions. Official HMMT archive, 2025. URL https://www.hmmt.org/www/archive/282. Accessed 2026-09-24.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Hugging Face. Open R1: A fully open reproduction of DeepSeek-R1. GitHub repository, 2025. URL https://github.com/huggingface/open-r1. Accessed 2026-09-24.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley, New York, 1976. ISBN 978-0-471-46510-2. URL https://books.google.com/books/about/Decisions_with_Multiple_Objectives.html?id=k_x9AAAAIAAJ.

Maurice G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023a. doi: 10.52202/075280-0943. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your LLMs capable of stable reasoning? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17594–17632, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.905. URL https://aclanthology.org/2025.findings-acl.905/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yi Liu. Two calls, two moments, and the vote-accuracy curve of repeated LLM inference. arXiv preprint arXiv:2605.03379, 2026. URL https://arxiv.org/abs/2605.03379.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. AceReason-Nemotron 1.1: Advancing math and code reasoning through SFT and RL synergy. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2506.13284. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/72398f80dc0f81159f79e6ff8fd1e376-Abstract-Conference.html.

Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior. arXiv preprint arXiv:2511.10661, 2025. URL https://arxiv.org/abs/2511.10661. Accepted to the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

R. Timothy Marler and Jasbir S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, 2004. doi: 10.1007/s00158-003-0368-6.

MathArena. AIME I 2024. Hugging Face dataset, 2025a. URL https://huggingface.co/datasets/MathArena/aime_2024_I. Competition problems; accessed 2026-09-24.

MathArena. AIME 2025. Hugging Face dataset, 2025b. URL https://huggingface.co/datasets/MathArena/aime_2025. Competition problems; accessed 2026-09-24.

MathArena. AIME II 2024. Hugging Face dataset, 2025c. URL https://huggingface.co/datasets/MathArena/aime_2024_II. Competition problems; accessed 2026-09-24.

MathArena. AIME 2026. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/MathArena/aime_2026. Competition problems; accessed 2026-09-24.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, MA, 1998. ISBN 978-0-7923-8278-2. doi: 10.1007/978-1-4615-5563-6.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model. arXiv preprint arXiv:2508.14444, 2025a. URL https://arxiv.org/abs/2508.14444.

NVIDIA. OpenReasoning-Nemotron-1.5B. Hugging Face model card, 2025b. URL https://huggingface.co/nvidia/OpenReasoning-Nemotron-1.5B. Accessed 2026-09-24.

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

OpenThoughts. OpenThinker3-1.5B. Hugging Face model card, 2025. URL https://huggingface.co/open-thoughts/OpenThinker3-1.5B. Accessed 2026-09-26.

Raydonal Ospina and Silvia L. P. Ferrari. Inflated beta distributions. *Statistical Papers*, 51(1): 111–126, 2010. doi: 10.1007/s00362-008-0125-4.

Allan Pinkus. *Totally Positive Matrices*. Number 181 in Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-19408-2. doi: 10.1017/CBO9780511691713. First published in print in 2009.

Nearchos Potamitis, Vansh Ramani, Har Ashish Arora, Dhairya Kuchhal, Lars Klein, and Akhil Arora. ReasonBENCH: Benchmarking the (In)Stability of LLM reasoning. arXiv preprint arXiv:2512.07795, 2025. URL https://arxiv.org/abs/2512.07795.

Qwen. Qwen3-30B-A3B-Thinking-2507. Hugging Face model card, 2025a. URL https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507. Accessed 2026-09-26.

Qwen. Qwen3-4B-Thinking-2507. Hugging Face model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507. Accessed 2026-09-26.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling. arXiv preprint arXiv:2411.17501, 2024. URL https://arxiv.org/abs/2411.17501.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979. ISBN 978-0-408-70929-3. URL https://www.dcs.gla.ac.uk/Keith/Preface.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-R1: Curriculum SFT, DPO and RL for long COT from scratch and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 318–327. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-industry.24. URL https://aclanthology.org/2025.acl-industry.24/.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. arXiv preprint arXiv:2506.14245, 2025b. URL https://arxiv.org/abs/2506.14245.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2406.12045. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b126cc38b8638e07bef37e7b2bb72bf-Abstract-Conference.html.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Second Conference on Language Modeling*, 2025. doi: 10.48550/arXiv.2502.03387. URL https://openreview.net/forum?id=T2TZ0RY4Zk.

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? arXiv preprint arXiv:2503.24235, 2025. URL https://arxiv.org/abs/2503.24235.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

CONTENTS

1 Introduction **1**

2 Shape of Success **2**

2.1 Response-bank Shape 2

2.2 The Discovery–Stability Tail Surface 3

2.3 Failure Modes as Missing Shape 3

3 TailPass@k: Uncertainty-aware Discovery–Stability Evaluation **3**

3.1 Bayesian Uncertainty over the Profile 4

3.2 Expected-Utility Scalarization: TailPass@k 4

3.3 Rubric-aware Categorical Evaluation 5

4 Experiments **5**

4.1 Response-bank Diagnostics 5

4.2 Uncertainty in TailPass@k & Utility Rankings 6

4.3 Prior Evidence from Greedy Decoding 7

4.4 Rubric-aware Categorical Evaluation 8

5 Related Work **9**

6 Conclusion: Strengths, Limitations & Future Directions **9**

A How to Read This Paper: From Failures to TailPass **19**

B Notation **19**

C TailPass@k Estimands, Estimators, & Posterior Inference **21**

C.1 Latent Estimands & Finite-sample Estimators 21

C.2 Profile Endpoints, Sensitivity, & Area 23

C.3 Closed-form Profile Covariance 24

C.4 Credible Intervals for Profile Metrics 24

C.5 Choosing a Prior for TailPass@k 25

C.6 Large-sample Limit 25

C.7 Rubric-aware Categorical TailPass@k 26

C.8 Utilities & Spectrum Summaries 28

D Simulations: Information and Model Selection **33**

D.1 Information beyond Accuracy and Endpoints 33

D.2 Model Selection with Matched Endpoints 34

D.3 Sensitivity Analysis 35

E Choosing Utilities and Reporting Budgets **35**
E.1 Match the utility to the consequence . . . 35
E.2 Elicit the exponent from a value judgment . . . 36
E.3 Set the budget from the operation and its cost . . . 36
E.4 Fix reporting choices before comparing models . . . 36

F Extended Related Work **37**

G Extended Experiment Analyses **38**
G.1 Convergence & Rank Correlation . . . 38
G.1.1 Matched Resampling . . . 38
G.1.2 Convergence . . . 39
G.1.3 Kendall's τ_b Rank Correlation . . . 39
G.2 $\mathrm{TailPass@}k$ Uncertainty . . . 40
G.3 Prior-Evidence . . . 45
G.3.1 Validation, Alignment, and Full-bank Sensitivity . . . 45
G.3.2 Budget Sensitivity of Greedy-prior Validation . . . 48
G.4 Categorical Evaluation . . . 49

H Evaluation Across Mathematics and Science **54**
H.1 Response banks and evaluation protocol . . . 54
H.2 Discovery, repeatability, and moment utilities . . . 54
H.3 Threshold-specific model choice on SuperGPQA . . . 56
H.4 Repeatability-weighted ranking on SuperGPQA . . . 58

I Empirical Response-bank Diagnostics for Scalar Repeated-Sampling Metrics **59**
I.1 Temperature Effects on the Discovery–Stability Spectrum . . . 61

J Experimental Setup & Reproducibility **65**
J.1 Models & Datasets . . . 65
J.2 Sampling Protocols . . . 66
J.3 Prior-Evidence Validation Protocol . . . 66
J.4 $\mathrm{TailPass@}8$ Uncertainty Protocol . . . 67
J.5 Response-bank Diagnostic Protocol . . . 67
J.6 Categorical Evaluation Protocol . . . 67
J.7 Reproducibility . . . 68
J.8 Token Statistics . . . 68

K Failure Modes of Repeated-Sampling Metrics: $\mathrm{pass@}k$ **68**
K.1 Ranking Instability of $\mathrm{pass@}k$ . . . 69
K.1.1 Algebraic View of the Budget Path . . . 70

K.1.2 Stability & Reversal Conditions 71
K.1.3 Numerical Checks 73
K.2 Large-Budget Non-Convergence of $\mathrm{pass@}k$ to Average Accuracy 74
K.2.1 What the Large-budget Regimes Converge to 74
K.2.2 $\mathrm{pass@}k$ Exceeds Average Accuracy for Nontrivial Success Probabilities . . 76
K.2.3 Explicit Counterexamples 77
K.2.4 Numerical Checks 78
K.2.5 Practical Interpretation 79
K.3 Structural Non-Identifiability of $\mathrm{pass@}k$ 80
K.3.1 Two Sources of Structural Non-Identifiability 80
K.3.2 Equivalence Classes & Identifiability Rank 82
K.3.3 Counting Bounds for Finite Banks 83
K.3.4 Exact Enumeration Experiments 84

L Auxiliary Assumption Failures **86**
L.1 Verifier & Selector Fragility of $\mathrm{pass@}k$ 87
L.2 Dependence & Effective Sample-Size Collapse 90
L.3 One-Sided Utility & Harmful-Attempt Blindness 91
L.4 Finite-bank Uncertainty & False Ranking 93

M Failure Modes of Threshold & Area Metrics for Repeated Sampling **95**
M.1 Large-budget Targets Depend on the Chosen Threshold 96
M.2 Threshold-induced Ranking Reversals 97
M.3 Dependence Can Inflate Stability Coordinates 98
M.4 Majority Voting Requires Answer-cluster Information 99
M.5 Area Metrics Hide Where the Area Comes from 99

A How to Read This Paper: From Failures to TailPass

The paper progresses from diagnosing failures of repeated-sampling metrics to developing TailPass and its utilities, followed by experiments and simulations. The reading order below connects these stages across the main text and appendices.

1. Diagnose the failure modes.

- Budget- and threshold-dependent ranking reversals (Appendices K.1 and M.2).
- Large-budget targets differing from accuracy (Appendices K.2 and M.1).
- Structural non-identifiability of response-bank shape (Appendix K.3).
- Verifier and selector fragility (Appendix L.1).
- Dependence and distorted discovery or stability probabilities (Appendices L.2 and M.3).
- Harmful-attempt blindness (Appendix L.3).
- Finite-bank uncertainty and false ranking (Appendix L.4).
- Missing answer-cluster information for modal-answer voting (Appendix M.4).
- Area summaries concealing threshold or budget structure (Appendix M.5).

2. Understand TailPass and its utilities. Read Sections 2–3 for (i) discovery–stability profiles; (ii) joint posterior uncertainty and auxiliary evidence; (iii) moment utilities that shift emphasis toward repeatability; and (iv) categorical rubrics. Appendix C gives estimators, inference, and nonlinear utilities; Appendix E connects utility and budget choices to operational requirements.

3. Assess the experimental and simulation evidence. Section 4 and Appendices G.1–G.4, H, and I cover mathematics and science evaluations, including ranking uncertainty, convergence, prior transfer, and rubric sensitivity; Appendix I.1 examines decoding temperature. Finish with Appendix D for information loss and threshold-specific model selection under matched accuracy and endpoints.

B Notation

The discovery–stability profile S records upper-tail probabilities of the correctness count. Linear scalar summaries U are expectations of utilities over that count, with threshold weights specifying the value of each additional correct sample. Posterior means, variances, and credible intervals use standard Bayesian conditioning notation (Gelman et al., 2013).

Symbol	Meaning
Response bank and binary profile	
M, N	Numbers of questions and observed attempts per question, indexed by i and j, respectively.
$R, R_{i,j}$	Binary response matrix $R \in \{0,1\}^{M\times N}$ and correctness indicator for attempt j on question i.
γ_i	Observed success count, $\sum_{j=1}^{N} R_{i,j}$.
$n, \gamma_i(n)$	Observed prefix length $n \le N$ and its success count for question i.
h_c	Number of questions with c observed successes, $\#\{i : \gamma_i = c\}$, for $c = 0, \dots, N$.
p_i, p	Latent single-attempt success probability for question i; p denotes a generic success probability.
k, t, ρ	Reporting budget, required success count $t \in \{1, \dots, k\}$, and fractional threshold $\rho \in [0, 1]$.
$X_{i,k}$	Correctness count in a fresh k-attempt evaluation, with $X_{i,k} \mid p_i \sim \mathrm{Binomial}(k, p_i)$.
$q_{k,t}(p)$	Binomial tail $\Pr\{\mathrm{Binomial}(k,p) \ge t\}$.
$S_{i,k,t}, S_{i,k}$	Question-level tail probability $q_{k,t}(p_i)$ and profile $(S_{i,k,1}, \dots, S_{i,k,k})$.
$S_{k,t}, S_k$	Dataset mean $M^{-1}\sum_i S_{i,k,t}$ and profile $(S_{k,1}, \dots, S_{k,k})$, denoted by TailPass@$k$.
$S_{k,1}, S_{k,k}$	Discovery and stability endpoints: probabilities of at least one success and all k successes, averaged over questions.

Symbol	Meaning
$S_k(\rho)$	Normalized profile $S_{k,\lceil k\rho \rceil}$ for $\rho > 0$; the discovery endpoint is $S_k(0^+) = S_{k,1}$.
$\widehat{S}^{\mathrm{fs}}_{i,k,t}, \widehat{S}^{\mathrm{fs}}_{k,t}$	Hypergeometric tail from $k \le N$ draws without replacement from row i, and its dataset average. Finite-bank $\mathrm{pass@}k = \widehat{S}^{\mathrm{fs}}_{k,1}$.
Posterior inference	
R^0, D, γ_i^0	Optional prior response bank, its trials per question, and prior success count. Without prior evidence, $D = \gamma_i^0 = 0$.
η	Prior-transfer weight, $0 \le \eta \le 1$.
α_i^0, β_i^0	Beta prior parameters; the prior-count construction uses $\alpha_i^0 = 1 + \eta\gamma_i^0$ and $\beta_i^0 = 1 + \eta(D - \gamma_i^0)$.
a_i, b_i	Beta posterior parameters $a_i = \alpha_i^0 + \gamma_i$ and $b_i = \beta_i^0 + N - \gamma_i$.
$\boldsymbol{\mu}_k, \mu_{k,t}$	Posterior mean $\mathbb{E}[S_k \mid R, R^0, \eta]$ and its t-th coordinate.
$\Sigma_k, \sigma^2_{k,t}$	Posterior covariance of S_k and coordinate variance $[\Sigma_k]_{t,t}$, conditional on (R, R^0, η).
$H, h, p_i^{(h)}$	Number of joint posterior draws, draw index, and sampled success probability; $S_k^{(h)}$ is the resulting profile draw.
δ, Q_q	Credible-interval tail mass and posterior quantile operator. An equal-tailed interval has endpoints $Q_{\delta/2}$ and $Q_{1-\delta/2}$, with posterior mass $1 - \delta$.
Scalar utilities	
$W, w(\rho)$	Probability measure over fractional thresholds and its density, when one exists.
$\omega_{t,k}, \boldsymbol{\omega}_k$	Nonnegative threshold weights summing to one and their column vector. A measure W assigns weight $\omega_{t,k} = W(C_{t,k})$ to $C_{1,k} = [0, 1/k]$ and $C_{t,k} = ((t-1)/k, t/k]$ for $t \ge 2$.
$U_\omega(k)$	Linear utility $\sum_t \omega_{t,k} S_{k,t}$; $U_W(k)$ denotes the utility induced by W.
$\lambda, \omega^{(\lambda)}_{t,k}$	Positive moment exponent and weights $(t/k)^\lambda - ((t-1)/k)^\lambda$.
$U_\lambda(k), U_\lambda@k$	Moment utility $\sum_t \omega^{(\lambda)}_{t,k} S_{k,t}$; both budget notations denote the same quantity.
$\mu_{\lambda,k}, \sigma^2_{\lambda,k}$	Posterior mean and variance of $U_\lambda(k)$.
$\mathrm{QRS}_W(k), L_2@k$	Question-first quadratic spectrum score $M^{-1}\sum_i (\sum_t \omega_{t,k} S^2_{i,k,t})^{1/2}$. Uniform weights give $L_2@k$.
Categorical outcomes	
$L, C_{i,j}, n_{i,c}$	Number of rubric categories, observed category label, and count of category c on question i; $c = 1, \ldots, L$.
$\pi_i, \boldsymbol{\alpha}_i$	Latent category-probability vector and its Dirichlet posterior parameters, $\alpha_{i,c} = 1 + \eta n^0_{i,c} + n_{i,c}$, where $n^0_{i,c}$ is the prior count.
$v, Y_{i,k}, A^v_{i,k}$	Category scores $v \in [0,1]^L$, future category-count vector $Y_{i,k} \mid \pi_i \sim \mathrm{Multinomial}(k, \pi_i)$, and mean bank score $A^v_{i,k} = v^\top Y_{i,k}/k$.
$T^v_{i,k}(\rho), T^v_k(\rho)$	Categorical tail $\Pr(A^v_{i,k} \ge \rho \mid \pi_i)$ and its dataset average.
$U_{v,\lambda}(k)$	Rubric moment utility $M^{-1}\sum_i \mathbb{E}[(A^v_{i,k})^\lambda \mid \pi_i]$.
Matched-endpoint simulation	
$F_m, s_{m,t}$	Model m's distribution of success probabilities across questions and population tail $s_{m,t} = \mathbb{E}_{p\sim F_m}[q_{16,t}(p)]$; $m \in \{A, B\}$.
w	Model B's total probability mass at $p = 0$ and $p = 1$, split equally; the remaining mass $1 - w$ is at $p = 1/2$.

C TailPass@k Estimands, Estimators, & Posterior Inference

Binary TailPass@k evaluation infers a posterior success probability for each question, maps those posteriors through the binomial tail profile, and evaluates any pre-specified threshold weights or nonlinear spectrum summaries on the same joint posterior draws. The profile remains the primary object; scalar summaries are declared functionals of that profile. Algorithm 1 summarizes the binary-response procedure, including optional prior evidence and uncertainty propagation. Algorithm 2 gives the categorical rubric procedure.

Algorithm 1 TailPass@k: Uncertainty-aware discovery–stability evaluation

Require: Response matrix $R \in \{0,1\}^{M\times N}$; optional prior matrix R^0; transfer weight η; reporting budget k; posterior draws H; optional pre-specified threshold weights $\boldsymbol{\omega}_k$ or weighting rule W.

for $i = 1, \dots, M$ **do**
  Compute $\gamma_i = \sum_j R_{i,j}$. If R^0 is unavailable, set $D = 0$ and $\gamma_i^0 = 0$; otherwise compute $\gamma_i^0 = \sum_d R^0_{i,d}$.
  Set $a_i = 1 + \eta\gamma_i^0 + \gamma_i$ and $b_i = 1 + \eta(D - \gamma_i^0) + N - \gamma_i$.
  Compute $\mu_{i,k,t}$ for $t = 1, \dots, k$ using the beta-ratio expression in Section 3.1.
end for
Average over questions to obtain $\boldsymbol{\mu}_k = (\mu_{k,1}, \dots, \mu_{k,k})$.
if a scalar linear utility is specified **then**
  Set finite-budget weights $\omega_{t,k} = \omega_{t,k}^{(W)}$ from the weighting rule W, or use the supplied vector $\boldsymbol{\omega}_k$ with $\omega_{t,k} \geq 0$ and $\sum_t \omega_{t,k} = 1$.
  Combine the mean-profile coordinates as $\mu_{\omega,k} = \sum_{t=1}^k \omega_{t,k}\mu_{k,t}$.
end if
Optionally compute Σ_k from Appendix C.3; if a linear utility is specified, compute $\sigma^2_{\omega,k} = \boldsymbol{\omega}_k^\top \Sigma_k \boldsymbol{\omega}_k$.
for $h = 1, \dots, H$ **do**
  Draw $p_i^{(h)} \sim \mathrm{Beta}(a_i, b_i)$ independently for all questions.
  Evaluate the joint profile draw $S_k^{(h)} = (S_{k,1}^{(h)}, \dots, S_{k,k}^{(h)})$.
  if a linear utility is specified **then**
    Compute $U_\omega^{(h)}(k) = \sum_{t=1}^k \omega_{t,k} S_{k,t}^{(h)}$.
  end if
  Evaluate any pre-specified nonlinear summaries on the same joint draw.
end for
Report $\boldsymbol{\mu}_k$, componentwise credible intervals, and any pre-specified weighted scalar summaries with posterior means and credible intervals.

C.1 Latent Estimands & Finite-sample Estimators

The latent profile is the estimand; the conventional finite-sample pass@k estimator is an observed-bank statistic. The basic unit is a fixed model–question pair. For a single evaluated model, the question index i absorbs the model label; in a multi-model comparison the same construction can be written with $p_{m,i}$ for model m and question i. The latent quantity p_i is the single-run probability of a correct response for that pair. It differs from the observed fraction γ_i/N, and correctness remains stochastic at the question level.

The observed row-count histogram has survival curve

$$\bar{H}_R(t) = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{\gamma_i \geq t\}, \qquad t = 1, \dots, N. \tag{5}$$

At $k = N$, this is exactly the finite-bank threshold profile: the full-bank count is fixed, so a threshold is either met or not met on each question.

For a prospective reporting budget k, the latent discovery estimand is

$$\text{Discovery}_i\ @k = \Pr(X_{i,k} \geq 1 \mid p_i) = 1 - (1 - p_i)^k,$$

where $X_{i,k} \mid p_i \sim \mathrm{Binomial}(k, p_i)$. The full threshold-profile estimand generalizes this endpoint:

$$S_{i,k,t}(p_i) = \Pr(X_{i,k} \geq t \mid p_i) = \sum_{r=t}^{k} \binom{k}{r} p_i^r (1-p_i)^{k-r}.$$

This is the probability that a fresh set of k independent attempts reaches threshold t under the latent success probability p_i.

The conventional finite-bank $\mathrm{pass@}k$ calculation used in code-generation evaluation (Chen et al., 2021) starts from a different object. Suppose N responses have already been generated for question i, γ_i of them are correct, and $N \geq k$. The question-level finite-bank $\mathrm{pass@}k$ value is

$$\mathrm{pass}_i\ @k = 1 - \frac{\binom{N-\gamma_i}{k}}{\binom{N}{k}},$$

with the convention that $\binom{N-\gamma_i}{k} = 0$ when $N - \gamma_i < k$. This value is the $t = 1$ case of the finite-bank threshold estimator below. It is a function of the observed finite sample, whereas $1 - (1-p_i)^k$ is the underlying latent target.

The same distinction extends to every threshold in the discovery–stability profile. Given $N \geq k$ observed responses and γ_i observed correct responses, the natural finite-sample estimator of the threshold coordinate is the hypergeometric tail

$$\widehat{S}^{\mathrm{fs}}_{i,k,t} = \sum_{r=t}^{k} \frac{\binom{\gamma_i}{r}\binom{N-\gamma_i}{k-r}}{\binom{N}{k}},$$

where terms with invalid binomial coefficients are interpreted as zero. This is the probability that a size-k subset drawn without replacement from the observed N responses contains at least t correct responses. When $N < k$, this without-replacement estimator is undefined, while the latent and Bayesian posterior-predictive profiles remain well defined for any prospective reporting budget k. For $t = 1$ and $N \geq k$ the estimator reduces to the conventional $\mathrm{pass@}k$ estimator above:

$$\widehat{S}^{\mathrm{fs}}_{i,k,1} = 1 - \frac{\binom{N-\gamma_i}{k}}{\binom{N}{k}}.$$

The finite-sample estimator and the latent estimand are linked but distinct. If the N observed responses are independent Bernoulli(p_i) draws and $N \geq k$, then $\widehat{S}^{\mathrm{fs}}_{i,k,t}$ is unbiased for $S_{i,k,t}(p_i)$. To see this, write the hypergeometric estimator as the average over all size-k subsets of the N observed responses of the indicator that the subset contains at least t correct responses. Each fixed subset has probability

$$\Pr(\mathrm{Binomial}(k, p_i) \geq t) = S_{i,k,t}(p_i)$$

of meeting the threshold, so the average has the same expectation. Its realized value can nevertheless vary substantially when N is small, because it depends on the particular observed sample. This finite-sample variability motivates reporting uncertainty alongside point estimates in LLM evaluation (Hariri et al., 2026c).

The Bayesian formulation in Section 3.1 keeps the latent target and the finite evidence separate. The observed count γ_i updates the posterior over p_i, and the reported posterior mean

$$\mu_{i,k,t} = \mathbb{E}\big[S_{i,k,t}(p_i) \mid R_i, R_i^0, \eta\big] = \Pr(X_{i,k} \geq t \mid R_i, R_i^0, \eta) = \sum_{r=t}^{k} \binom{k}{r} \frac{\mathrm{B}(a_i + r, b_i + k - r)}{\mathrm{B}(a_i, b_i)}$$

is the beta-binomial posterior predictive tail for a fresh budget of k independent attempts. This quantity differs from substituting γ_i/N for p_i and from the without-replacement finite-sample estimator $\widehat{S}^{\mathrm{fs}}_{i,k,t}$, although all three target the same operational event in different ways. Appendix C.3 gives the covariance formula that quantifies posterior uncertainty in the latent profile $S_{k,t}$.

At the dataset level, the latent estimand is the expected fraction of questions that meet threshold t under fresh sampling:

$$S_{k,t}(p_{1:M}) = \frac{1}{M} \sum_{i=1}^{M} S_{i,k,t}(p_i).$$

The corresponding finite-sample estimator averages the question-level hypergeometric estimates,

$$\widehat{S}^{\mathrm{fs}}_{k,t} = \frac{1}{M}\sum_{i=1}^{M} \widehat{S}^{\mathrm{fs}}_{i,k,t}. \tag{6}$$

For finite-budget utility weights, the corresponding finite-sample scalar is

$$\widehat{U}^{\mathrm{fs}}_{\omega}(k) = \sum_{t=1}^{k} \omega_{t,k}\widehat{S}^{\mathrm{fs}}_{k,t}.$$

This is the same linear scalarization as $U_\omega(k)$ applied to the observed finite-sample profile. Bayesian posterior summaries quantify uncertainty in the latent conditional mean. A realized future benchmark with k attempts per question adds a sampling layer:

$$Z_{i,k,t} := \mathbf{1}\{X_{i,k} \geq t\}, \qquad \widetilde{S}_{k,t} := \frac{1}{M}\sum_{i=1}^{M} Z_{i,k,t}.$$

Then

$$\mathbb{E}\Big[\widetilde{S}_{k,t} \mid p_{1:M}\Big] = S_{k,t}(p_{1:M}),$$

so $S_{k,t}$ is the conditional mean of the realized finite-sample benchmark fraction. The realized fraction $\widetilde{S}_{k,t}$ contains an additional future-sampling layer; conditional on $p_{1:M}$ its variance is

$$\mathrm{Var}\Big(\widetilde{S}_{k,t} \mid p_{1:M}\Big) = \frac{1}{M^2}\sum_{i=1}^{M} S_{i,k,t}(p_i)\left(1 - S_{i,k,t}(p_i)\right),$$

assuming independent future samples across questions.

A posterior predictive analysis of a new realized benchmark would additionally include the Bernoulli threshold indicators $Z_{i,k,t}$.

C.2 Profile Endpoints, Sensitivity, & Area

The discovery and strict-stability endpoints are

$$S_{k,1} = \frac{1}{M}\sum_{i}\{1 - (1-p_i)^k\}, \tag{7}$$

$$S_{k,k} = \frac{1}{M}\sum_{i} p_i^k. \tag{8}$$

Intermediate thresholds represent partial repeatability. The sensitivity of each coordinate to the per-question success probability follows from differentiating the binomial tail:

$$\frac{\partial}{\partial p} q_{k,t}(p) = k\binom{k-1}{t-1} p^{t-1}(1-p)^{k-t}. \tag{9}$$

This is the density of a $\mathrm{Beta}(t, k-t+1)$ distribution. For $k > 1$, discovery is most sensitive near $p = 0$, strict stability near $p = 1$, and middle thresholds near intermediate success probabilities. At $k = 1$, discovery and stability coincide and the derivative is constant.

The tail-sum formula for nonnegative integer-valued random variables (Billingsley, 1995) gives

$$\frac{1}{k}\sum_{t=1}^{k} q_{k,t}(p) = \frac{1}{k}\mathbb{E}[X_{i,k} \mid p] = p, \qquad \frac{1}{k}\sum_{t=1}^{k} S_{k,t} = \frac{1}{M}\sum_{i=1}^{M} p_i. \tag{10}$$

Thus profiles with the same area have the same average single-sample accuracy, while their allocation across discovery and repeatability thresholds can differ.

C.3 Closed-form Profile Covariance

For fixed k and thresholds $t, s \in \{1, \dots, k\}$, define the question-level mixed posterior moment

$$\omega_{i,k,t,s} := \mathbb{E}\big[S_{i,k,t}(p_i) S_{i,k,s}(p_i) \mid R_i, R_i^0, \eta\big] .$$

Expanding both tail probabilities gives

$$\omega_{i,k,t,s} = \sum_{r=t}^{k} \sum_{\ell=s}^{k} \binom{k}{r} \binom{k}{\ell} \frac{\mathrm{B}(a_i + r + \ell, b_i + 2k - r - \ell)}{\mathrm{B}(a_i, b_i)}.$$

Because the posterior factorizes across questions, the covariance between two dataset-level profile coordinates is

$$[\Sigma_k]_{t,s} = \mathrm{Cov}\big(S_{k,t}, S_{k,s} \mid R, R^0, \eta\big) = \frac{1}{M^2} \sum_{i=1}^{M} \left(\omega_{i,k,t,s} - \mu_{i,k,t}\mu_{i,k,s}\right). \tag{11}$$

The diagonal case $t = s$ gives the marginal posterior variance of the threshold coordinate. Writing

$$\nu_{i,k,t} := \omega_{i,k,t,t} = \mathbb{E}\big[S_{i,k,t}(p_i)^2 \mid R_i, R_i^0, \eta\big] ,$$

the coordinate variance is

$$\sigma_{k,t}^2 := [\Sigma_k]_{t,t} = \frac{1}{M^2} \sum_{i=1}^{M} \left(\nu_{i,k,t} - \mu_{i,k,t}^2\right).$$

Posterior simulation estimates the same covariance by evaluating all thresholds on each shared draw of $p_{1:M}$. For

$$S_k^{(h)} = \left(S_{k,1}^{(h)}, \dots, S_{k,k}^{(h)}\right), \qquad \widehat{\mu}_{k,t} = \frac{1}{H} \sum_{h=1}^{H} S_{k,t}^{(h)},$$

the joint-draw covariance estimator is

$$[\widehat{\Sigma}_k]_{t,s} = \frac{1}{H-1} \sum_{h=1}^{H} \left(S_{k,t}^{(h)} - \widehat{\mu}_{k,t}\right) \left(S_{k,s}^{(h)} - \widehat{\mu}_{k,s}\right), \qquad t, s = 1, \dots, k.$$

Using the same latent posterior draw across all thresholds preserves the posterior dependence among profile coordinates.

C.4 Credible Intervals for Profile Metrics

All uncertainty intervals in this Bayesian formulation are posterior credible intervals: they summarize uncertainty in a posterior random variable after conditioning on the observed response matrices and the prior specification (Gelman et al., 2013). A frequentist confidence interval would instead be defined by repeated-sampling coverage over hypothetical datasets.

For a fixed budget k and threshold t, write

$$\mathcal{S}_{k,t} := S_{k,t} \mid R, R^0, \eta.$$

The componentwise uncertainty summary is the equal-tailed posterior credible interval

$$\left[Q_{\delta/2}(\mathcal{S}_{k,t}), Q_{1-\delta/2}(\mathcal{S}_{k,t})\right], \tag{12}$$

where Q_q is the q-th posterior quantile. The interval is defined by the posterior distribution itself and can be estimated from joint posterior draws of the profile. Posterior quantiles retain skewness and boundary effects without a Gaussian approximation (Gelman et al., 2013).

A moment-based alternative is to combine the exact posterior mean and variance with a normal approximation:

$$\left[\mu_{k,t} - z_{1-\delta/2}\sigma_{k,t}, \mu_{k,t} + z_{1-\delta/2}\sigma_{k,t}\right],$$

where $z_{1-\delta/2}$ is the corresponding standard-normal quantile and $\sigma_{k,t}^2 = [\Sigma_k]_{t,t}$. Because $0 \le S_{k,t} \le 1$, this interval is often clipped to $[0, 1]$ in applications. Moment-based Bayesian summaries are

computationally convenient and have been used in Bayesian LLM evaluation (Hariri et al., 2026c); the normal approximation requires only the first two posterior moments, but it determines posterior quantiles accurately only when the posterior distribution of $\mathcal{S}_{k,t}$ is well approximated by a Gaussian distribution (Gelman et al., 2013).

Another Bayesian choice is a highest posterior density region, defined as a posterior set with mass $1-\delta$ that contains points of highest posterior density. Such regions can be disconnected for multimodal posteriors (Gelman et al., 2013). For these profile metrics, equal-tailed posterior quantiles provide a direct summary of componentwise uncertainty. Closed-form means and covariances give exact posterior moments against which Monte Carlo estimates can be checked. Each equal-tailed interval has posterior coverage $1-\delta$ for its coordinate; simultaneous coverage across the entire profile requires a joint credible-set construction.

C.5 Choosing a Prior for TailPass@k

TailPass uses one prior on p_i for the full profile and all its utilities. Write the binary prior as $\mathrm{Beta}(\nu_i m_i, \nu_i(1-m_i))$, with mean $m_i \in (0,1)$ and concentration $\nu_i > 0$. After n observed attempts, the prior receives weight $\nu_i/(\nu_i+n)$ in the posterior mean of p_i (Gelman et al., 2013). Without auxiliary evidence, the default $m_i = 1/2$, $\nu_i = 2$ has a direct count-space interpretation: Equation (4) with no observations gives $\Pr(X_{i,k} = r) = 1/(k+1)$ for every $r = 0, \dots, k$. With a disjoint auxiliary bank, the construction in Section 3.1 instead gives $\nu_i = 2 + \eta D$ and $m_i = (1 + \eta\gamma_i^0)/(2+\eta D)$, so ηD specifies the additional prior evidence.

When calibration data are available, prior settings ϑ, such as η or the baseline mean and concentration, can be selected by predicting the whole success-count distribution. On calibration questions separate from the final benchmark, fit each candidate using n attempts per question, matching the intended observation budget. Let $\mu^{\vartheta}_{i,k,t}$ be its posterior mean tail and let $Z_{i,k,t} = \mathbf{1}\{X_{i,k} \geq t\}$ come from k held-out attempts under the target sampling protocol. Minimize the ranked probability score (Gneiting & Raftery, 2007), averaged over the M calibration questions:

$$\mathcal{L}_k(\vartheta) = \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{k}\big(\mu^{\vartheta}_{i,k,t} - Z_{i,k,t}\big)^2. \tag{13}$$

The selected prior rule is then fixed for the benchmark and its utility views.

The score estimates profile error up to an additive term shared by all priors. Conditional on the fitting data and $p_{1:M}$, independence of the held-out attempts gives, for each fixed candidate ϑ,

$$\mathbb{E}[\mathcal{L}_k(\vartheta)] = \mathcal{E}_k(\vartheta) + \frac{1}{M}\sum_{i,t} q_{k,t}(p_i)\{1-q_{k,t}(p_i)\}, \qquad \mathcal{E}_k(\vartheta) := \frac{1}{M}\sum_{i,t}\big(\mu^{\vartheta}_{i,k,t} - q_{k,t}(p_i)\big)^2. \tag{14}$$

The second term is independent of the prior choice. For the same questions, let $\widehat{U}^{\vartheta}_{\omega}(k) = M^{-1}\sum_{i,t}\omega_{t,k}\mu^{\vartheta}_{i,k,t}$. Every nonnegative, normalized threshold weighting obeys

$$\big|\widehat{U}^{\vartheta}_{\omega}(k) - U_{\omega}(k)\big|^2 \leq \mathcal{E}_k(\vartheta). \tag{15}$$

Indeed, Jensen's inequality bounds the left side by the average squared coordinate error weighted by $\omega_{t,k}$, which is at most the unweighted sum. Thus profile prediction supplies a common error criterion for discovery, stability, and every $U_\lambda(k)$.

For greedy evidence, $\eta \in \{0, .25, .5, 1\}$ provides a concrete candidate set. At $n = 4$, full transfer reduces aggregate held-out MAE for every tested $U_\lambda(16)$, $\lambda \in \{.5, 1, 2\}$ (Table 3), benefiting several operating utilities in the same low-sample regime.

For categorical rubrics, the analogous parameterization is $\mathrm{Dirichlet}(\nu_i \boldsymbol{m}_i)$, with $\boldsymbol{m}_i$ a positive probability vector. Splitting a category's prior mass among refined labels preserves the original prior after merging, by Dirichlet aggregation (Gelman et al., 2013). Finer rubric distinctions can therefore retain the same prior evidence on coarse outcomes.

C.6 Large-sample Limit

For a fixed question i, conditional on p_i, the fraction $X_{i,k}/k$ is the sample mean of k independent Bernoulli(p_i) variables. The weak law of large numbers implies that $X_{i,k}/k$ converges in probability

to p_i as $k \to \infty$ (Billingsley, 1995). Therefore, for every $\rho \in (0, 1]$ with $\rho \neq p_i$,

$$q_{k,\lceil k\rho \rceil}(p_i) \longrightarrow \mathbf{1}\{p_i > \rho\}. \tag{16}$$

At continuity points of the empirical distribution of p_i, averaging over questions gives

$$S_k(\rho) \longrightarrow \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{p_i > \rho\}. \tag{17}$$

Equivalently, the question-level limit away from $\rho = p_i$ is

$$S_{i,k}(\rho) = \Pr\left(\frac{X_{i,k}}{k} \geq \rho \mid p_i\right) \longrightarrow S_{i,\infty}(\rho) := \mathbf{1}\{p_i \geq \rho\},$$

where the value at $\rho = p_i$ is fixed by convention and does not affect any integral identity. Thus the limiting profile is the survival function of the per-question success probability. In particular,

$$S_{i,\infty}(0^+) = \mathbf{1}\{p_i > 0\}$$

is ideal discovery, while

$$S_{i,\infty}(1) = \mathbf{1}\{p_i = 1\}$$

is ideal stability. The area under the limiting curve is again the single-sample success probability:

$$\int_0^1 S_{i,\infty}(\rho)\, d\rho = p_i.$$

C.7 Rubric-aware Categorical TailPass@k

Rubric-aware TailPass replaces the Bernoulli success variable with a finite categorical outcome, using the Dirichlet–multinomial conjugate model (Gelman et al., 2013; Hariri et al., 2026c). Let $C_{i,j} \in \{1, \ldots, L\}$ denote the category assigned to attempt j on question i by a fixed rubric schema, and let $n_{i,c} = \sum_{j=1}^{N} \mathbf{1}\{C_{i,j} = c\}$ be the observed category counts. Optional prior evidence C^0 gives counts $n^0_{i,c}$ over D prior attempts. With transfer weight $\eta \in [0, 1]$, use the Dirichlet prior

$$\pi_i \sim \text{Dirichlet}(\boldsymbol{\alpha}^0_i), \qquad \alpha^0_{i,c} = 1 + \eta n^0_{i,c},$$

where $n^0_{i,c} = 0$ when no prior response bank is supplied. Categorical conjugacy gives

$$\pi_i \mid C_i, C^0_i, \eta \sim \text{Dirichlet}(\boldsymbol{\alpha}_i), \qquad \alpha_{i,c} = 1 + \eta n^0_{i,c} + n_{i,c}. \tag{18}$$

For $L = 2$ with labels ordered as wrong or correct, $p_i = \pi_{i,2}$ has the Beta–Bernoulli posterior in Equation (3), with $a_i = \alpha_{i,2}$ and $b_i = \alpha_{i,1}$.

For a prospective budget k, let

$$\mathcal{Y}_{k,L} = \left\{ y \in \mathbb{N}_0^L : \sum_{c=1}^{L} y_c = k \right\}$$

be the set of possible category-count vectors. Conditional on π_i,

$$Y_{i,k} \mid \pi_i \sim \text{Multinomial}(k, \pi_i), \qquad m_k(y; \pi_i) = \frac{k!}{\prod_{c=1}^{L} y_c!} \prod_{c=1}^{L} \pi_{i,c}^{y_c}.$$

The posterior predictive distribution of $Y_{i,k}$ is Dirichlet–multinomial:

$$d_{i,k}(y) := \Pr(Y_{i,k} = y \mid C_i, C^0_i, \eta) = \frac{k!}{\prod_{c=1}^{L} y_c!} \frac{\mathrm{B}_L(\boldsymbol{\alpha}_i + y)}{\mathrm{B}_L(\boldsymbol{\alpha}_i)}, \tag{19}$$

where $\mathrm{B}_L(a) = \prod_{c=1}^{L} \Gamma(a_c) / \Gamma(\sum_c a_c)$ is the multivariate beta function.

Let $v \in [0, 1]^L$ be the rubric score vector and define the normalized score of a future k-attempt bank by

$$A^v_{i,k} = \frac{1}{k} v^\top Y_{i,k}.$$

The question-level categorical tail profile is

$$T^v_{i,k}(\rho;\pi_i) = \Pr(A^v_{i,k} \geq \rho \mid \pi_i) = \sum_{y\in\mathcal{Y}_{k,L}} \mathbf{1}\{v^\top y/k \geq \rho\} m_k(y;\pi_i), \qquad \rho \in [0,1]. \tag{20}$$

The dataset-level profile is $T^v_k(\rho) = M^{-1}\sum_i T^v_{i,k}(\rho;\pi_i)$. On the finite grid $\rho_t = t/k$, $t = 1,\dots,k$, write $T^v_{i,k,t} = T^v_{i,k}(t/k;\pi_i)$ and $T^v_{k,t} = M^{-1}\sum_i T^v_{i,k,t}$. The posterior mean of the grid profile is

$$\mu^v_{k,t} = \frac{1}{M}\sum_{i=1}^{M}\sum_{y\in\mathcal{Y}_{k,L}} \mathbf{1}\{v^\top y \geq t\} d_{i,k}(y). \tag{21}$$

The binary profile is recovered by choosing $L = 2$ with wrong or correct ordering, $v = (0,1)$, and identifying y_2 with the number of correct responses.

The covariance of two grid coordinates follows from Dirichlet monomial moments. Let $c_k(y) = k!/\prod_c y_c!$ and $g^v_t(y) = \mathbf{1}\{v^\top y \geq t\}$. The question-level second mixed moment is

$$\Omega^v_{i,k,t,s} = \sum_{y\in\mathcal{Y}_{k,L}}\sum_{z\in\mathcal{Y}_{k,L}} g^v_t(y) g^v_s(z) c_k(y) c_k(z) \frac{\mathrm{B}_L(\boldsymbol{\alpha}_i + y + z)}{\mathrm{B}_L(\boldsymbol{\alpha}_i)}. \tag{22}$$

Writing $\mu^v_{i,k,t} = \sum_y g^v_t(y) d_{i,k}(y)$, posterior independence across questions gives

$$[\Sigma^v_k]_{t,s} = \frac{1}{M^2}\sum_{i=1}^{M}\left(\Omega^v_{i,k,t,s} - \mu^v_{i,k,t}\mu^v_{i,k,s}\right). \tag{23}$$

Equations (21)–(23) are exact for fixed k and finite L. In practice, the same quantities can be estimated by joint posterior draws $\pi_i^{(h)} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_i)$, evaluating $T^v_{i,k}(\rho;\pi_i^{(h)})$ for all thresholds, and averaging over questions. For a fixed category map, shared draws preserve posterior dependence among thresholds, budgets, scalar summaries, and alternative score vectors. Comparisons across different category maps require a joint model, such as a common refinement of their labels, to define cross-rubric posterior covariance.

Moment-family rubric utilities. For a probability measure W on $[0,1]$, a linear threshold utility is

$$U_{i,v,W}(k;\pi_i) = \int_0^1 T^v_{i,k}(\rho;\pi_i)\, dW(\rho).$$

The moment family in Appendix C.8 uses $dW_\lambda(\rho) = \lambda\rho^{\lambda-1}d\rho$ for $\lambda > 0$. Since $0 \leq A^v_{i,k} \leq 1$,

$$U_{i,v,\lambda}(k;\pi_i) = \int_0^1 T^v_{i,k}(\rho;\pi_i)\lambda\rho^{\lambda-1}d\rho = \mathbb{E}\left[(A^v_{i,k})^\lambda \mid \pi_i\right] = \sum_{y\in\mathcal{Y}_{k,L}}\left(\frac{v^\top y}{k}\right)^\lambda m_k(y;\pi_i). \tag{24}$$

The posterior mean and variance are

$$\mu_{v,\lambda,k} = \frac{1}{M}\sum_{i=1}^{M}\sum_{y\in\mathcal{Y}_{k,L}}\left(\frac{v^\top y}{k}\right)^\lambda d_{i,k}(y), \tag{25}$$

$$\sigma^2_{v,\lambda,k} = \frac{1}{M^2}\sum_{i=1}^{M}\left(\Xi_{i,v,\lambda,k} - \mu^2_{i,v,\lambda,k}\right), \tag{26}$$

where

$$\mu_{i,v,\lambda,k} = \sum_y \left(\frac{v^\top y}{k}\right)^\lambda d_{i,k}(y)$$

and

$$\Xi_{i,v,\lambda,k} = \sum_{y\in\mathcal{Y}_{k,L}}\sum_{z\in\mathcal{Y}_{k,L}}\left(\frac{v^\top y}{k}\right)^\lambda\left(\frac{v^\top z}{k}\right)^\lambda c_k(y)c_k(z)\frac{\mathrm{B}_L(\boldsymbol{\alpha}_i + y + z)}{\mathrm{B}_L(\boldsymbol{\alpha}_i)}.$$

Algorithm 2 Rubric-aware categorical TailPass@k

Require: Categorical response matrix $C \in \{1, \dots, L\}^{M \times N}$; optional prior matrix C^0; transfer weight η; score vector $v \in [0,1]^L$; reporting budget k; posterior draws H; optional threshold set $\mathcal{R} \subset [0,1]$ and moment exponents Λ.

if $\mathcal{R}$ is unavailable **then**
  Set $\mathcal{R} = \{1/k, 2/k, \dots, 1\}$.
end if
for $i = 1, \dots, M$ **do**
  Compute $n_{i,c} = \sum_j \mathbf{1}\{C_{i,j} = c\}$ for $c = 1, \dots, L$.
  If C^0 is unavailable, set $D = 0$ and $n^0_{i,c} = 0$; otherwise compute $n^0_{i,c} = \sum_d \mathbf{1}\{C^0_{i,d} = c\}$.
  Set $\alpha_{i,c} = 1 + \eta n^0_{i,c} + n_{i,c}$ for $c = 1, \dots, L$.
  Compute $d_{i,k}(y)$ on $\mathcal{Y}_{k,L}$ using Equation (19).
  For each $\rho \in \mathcal{R}$, compute $\mu^v_{i,k}(\rho) = \sum_{y \in \mathcal{Y}_{k,L}} \mathbf{1}\{v^\top y/k \geq \rho\} d_{i,k}(y)$.
end for
Average over questions to obtain $\mu^v_k(\rho) = M^{-1} \sum_i \mu^v_{i,k}(\rho)$ for all $\rho \in \mathcal{R}$.
Optionally compute Σ^v_k on the canonical grid $\{1/k, \dots, 1\}$ using Equation (23).
if moment exponents are specified **then**
  Compute posterior mean and variance for each $U_{v,\lambda}(k)$, $\lambda \in \Lambda$, using Equations (25)–(26).
end if
for $h = 1, \dots, H$ **do**
  Draw $\pi^{(h)}_i \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}_i)$ independently for all questions.
  Evaluate $T^{v,(h)}_k(\rho) = M^{-1} \sum_i T^v_{i,k}(\rho; \pi^{(h)}_i)$ for all $\rho \in \mathcal{R}$.
  if moment exponents are specified **then**
    Compute $U^{(h)}_{v,\lambda}(k) = M^{-1} \sum_i U_{i,v,\lambda}(k; \pi^{(h)}_i)$ for each $\lambda \in \Lambda$.
  end if
end for
Report $\mu^v_k(\rho)$, componentwise credible intervals, and any pre-specified rubric utilities with posterior means and credible intervals.

Equal-tailed credible intervals are obtained by evaluating $M^{-1} \sum_i U_{i,v,\lambda}(k; \pi^{(h)}_i)$ on the same joint posterior draws. The limiting case $\lambda \downarrow 0$ gives the rubric-discovery endpoint

$$U_{i,v,0^+}(k; \pi_i) = \Pr(A^v_{i,k} > 0 \mid \pi_i) = 1 - \left(\sum_{c: v_c = 0} \pi_{i,c} \right)^k,$$

with the convention that only categories receiving positive rubric score count as discovered.

For integer moment parameters, the expected utility can often be evaluated without enumerating $\mathcal{Y}_{k,L}$. Let ξ^v_i be the score of one future attempt and let $m_{i,r} = \mathbb{E}[(\xi^v_i)^r \mid \pi_i] = \sum_c \pi_{i,c} v^r_c$. Then

$$\mathbb{E}[A^v_{i,k} \mid \pi_i] = m_{i,1}, \qquad \mathbb{E}[(A^v_{i,k})^2 \mid \pi_i] = \frac{k m_{i,2} + k(k-1) m^2_{i,1}}{k^2},$$

and

$$\mathbb{E}[(A^v_{i,k})^4 \mid \pi_i] = \frac{1}{k^4} \{ k m_{i,4} + 4k(k-1) m_{i,3} m_{i,1} + 3k(k-1) m^2_{i,2} + 6k(k-1)(k-2) m_{i,2} m^2_{i,1} + k(k-1)(k-2)(k-3) m^4_{i,1} \},$$

where terms with negative factors are omitted when $k < 4$. The categorical experiments use these raw-moment identities for $\lambda \in \{1, 2, 4\}$ and the closed-form 0^+ discovery endpoint above.

C.8 Utilities & Spectrum Summaries

The discovery–stability profile is a multi-criteria object whose coordinates are indexed by correctness thresholds. This matters for agent evaluation: pass@k and repeated-sampling coverage measure whether at least one sampled solution succeeds (Chen et al., 2021; Brown et al., 2024), whereas pass^k requires every attempt to succeed (Yao et al., 2025) and G-Pass@k_τ varies the required fraction of

correct attempts (Liu et al., 2025). The stability endpoint and intermediate thresholds in the profile capture the same potential-to-stability continuum. A scalar combination is therefore a preference over operational regimes. The scalar should be monotone in every profile coordinate, use a pre-specified weighting rule, preserve the interpretation of a dataset score as an average over questions, and admit posterior uncertainty propagation from the same joint draws used for the profile.

Normalized thresholds & finite-budget weights. Fractional thresholds re-index the finite profile on the normalized correctness scale. For $\rho \in [0,1]$, define

$$S_{i,k}(\rho) = \Pr\left(\frac{X_{i,k}}{k} \geq \rho \mid p_i\right),$$

with $S_{i,k}(\rho) = S_{i,k,\lceil k\rho \rceil}$ for $\rho \in (0,1]$ and with the right limit $S_{i,k}(0^+)$ used at the discovery endpoint. Let W be a probability measure on $[0,1]$. At finite budget k, only the mass assigned to the intervals

$$C_{1,k} = [0, 1/k], \qquad C_{t,k} = ((t-1)/k, t/k] \quad (t \geq 2)$$

affects the score. The induced finite-budget weights are

$$\omega_{t,k}^{(W)} = W(C_{t,k}), \qquad t = 1, \ldots, k.$$

Conversely, any nonnegative vector $\boldsymbol{\omega}_k = (\omega_{1,k}, \ldots, \omega_{k,k})^\top$ with $\sum_t \omega_{t,k} = 1$ is an admissible finite-budget viewpoint.

Linear expected-utility combinations. The most direct scalarization is the linear score

$$U_{i,\omega}(k) = \sum_{t=1}^{k} \omega_{t,k} S_{i,k,t}, \qquad \omega_{t,k} \geq 0, \quad \sum_{t=1}^{k} \omega_{t,k} = 1.$$

Equivalently, $U_{i,\omega}(k) = \mathbb{E}[u_{\omega,k}(X_{i,k}) \mid p_i]$ for the nondecreasing utility $u_{\omega,k}(x) = \sum_{t=1}^{x} \omega_{t,k}$. This is a weighted-sum scalarization of the threshold probabilities (Marler & Arora, 2004), with weights fixed by the evaluator's trade-offs. For discovery–stability evaluation, $\omega_{t,k}$ has a direct operational interpretation as the marginal value of the t-th correct sample. The score commutes with the dataset average,

$$U_\omega(k) = \frac{1}{M} \sum_{i=1}^{M} U_{i,\omega}(k) = \sum_{t=1}^{k} \omega_{t,k} S_{k,t}.$$

Uniform weights recover accuracy, endpoint weights recover discovery and stability, and the moment-family weights below shift mass from discovery-oriented thresholds toward stability-oriented thresholds.

When evaluation is reported from an observed response bank with $N \geq k$, the same linear scalarization can be applied to the finite-sample profile:

$$\widehat{U}_\omega^{\mathrm{fs}}(k) := \sum_{t=1}^{k} \omega_{t,k} \widehat{S}_{k,t}^{\mathrm{fs}}.$$

The posterior mean and variance of $U_\omega(k)$ are obtained by applying the same linear functional to the posterior mean profile and covariance matrix:

$$\begin{aligned} \mu_{\omega,k} &:= \mathbb{E}\big[U_\omega(k) \mid R, R^0, \eta\big] = \sum_{t=1}^{k} \omega_{t,k} \mu_{k,t}, \\ \sigma_{\omega,k}^2 &:= \mathrm{Var}\big(U_\omega(k) \mid R, R^0, \eta\big) = \boldsymbol{\omega}_k^\top \Sigma_k \boldsymbol{\omega}_k, \qquad \boldsymbol{\omega}_k = (\omega_{1,k}, \ldots, \omega_{k,k})^\top. \end{aligned} \tag{27}$$

Equivalently,

$$\sigma_{\omega,k}^2 = \sum_{t=1}^{k} \omega_{t,k}^2 \sigma_{k,t}^2 + 2 \sum_{1 \leq t < s \leq k} \omega_{t,k} \omega_{s,k} [\Sigma_k]_{t,s}.$$

Equal-tailed credible intervals are estimated from the scalar posterior draws

$$U_\omega^{(h)}(k) = \sum_{t=1}^{k} \omega_{t,k} S_{k,t}^{(h)}, \qquad h = 1, \ldots, H. \tag{28}$$

Weight families. Endpoint and uniform weights give the simplest baselines:

$$\omega_{t,k}^{\mathrm{disc}} = \mathbf{1}\{t=1\}, \qquad \omega_{t,k}^{\mathrm{stab}} = \mathbf{1}\{t=k\}, \qquad \omega_{t,k}^{\mathrm{acc}} = \frac{1}{k}.$$

The moment family is a one-parameter threshold measure. For $\lambda > 0$, let W_λ have density $w_\lambda(\rho) = \lambda\rho^{\lambda-1}$. Then

$$U_{i,W_\lambda}(k) = \int_0^1 \lambda\rho^{\lambda-1} S_{i,k}(\rho)\, d\rho = \mathbb{E}\left[\left(\frac{X_{i,k}}{k}\right)^\lambda \,\middle|\, p_i\right],$$

with induced finite-budget weights

$$\omega_{t,k}^{(\lambda)} = \left(\frac{t}{k}\right)^\lambda - \left(\frac{t-1}{k}\right)^\lambda, \qquad t = 1, \dots, k. \tag{29}$$

Small λ concentrates mass near discovery thresholds, large λ concentrates mass near stability thresholds, and $\lambda = 1$ gives uniform threshold coverage. The posterior moments of the moment-family utility are

$$\begin{aligned} \mu_{\lambda,k} &= \sum_{t=1}^{k} \omega_{t,k}^{(\lambda)} \mu_{k,t}, \\ \sigma_{\lambda,k}^2 &= \boldsymbol{\omega}_{\lambda,k}^\top \Sigma_k \boldsymbol{\omega}_{\lambda,k}, \qquad \sigma_{\lambda,k} = \sqrt{\sigma_{\lambda,k}^2}. \end{aligned} \tag{30}$$

where $\boldsymbol{\omega}_{\lambda,k} = (\omega_{1,k}^{(\lambda)}, \dots, \omega_{k,k}^{(\lambda)})^\top$.

The maximum-entropy principle selects a distribution by maximizing entropy subject to the available constraints (Jaynes, 1957). Let $\Theta \in [0,1]$ denote the operational threshold and suppose only its mean $\bar\rho = \mathbb{E}[\Theta]$, with $0 < \bar\rho < 1$, is specified. Relative to Lebesgue measure on $[0,1]$, the entropy-maximizing density is

$$w_\tau^{\mathrm{ME}}(\rho) = \frac{\exp(\tau\rho)}{Z(\tau)}, \qquad Z(\tau) = \begin{cases} \dfrac{e^\tau - 1}{\tau}, & \tau \neq 0, \\ 1, & \tau = 0. \end{cases} \tag{31}$$

For any density w with the same mean and finite differential entropy $H(w) = -\int_0^1 w(\rho)\log w(\rho)\, d\rho$,

$$H(w_\tau^{\mathrm{ME}}) - H(w) = \int_0^1 w(\rho) \log \frac{w(\rho)}{w_\tau^{\mathrm{ME}}(\rho)}\, d\rho \geq 0,$$

which establishes the maximizing property; densities with entropy $-\infty$ cannot improve it. The natural parameter $\tau = \tau(\bar\rho)$ is chosen so that

$$\bar\rho = \begin{cases} \dfrac{e^\tau(\tau-1)+1}{\tau(e^\tau - 1)}, & \tau \neq 0, \\ \dfrac{1}{2}, & \tau = 0. \end{cases}$$

The corresponding finite-budget weights are

$$\omega_{t,k}^{\mathrm{ME}}(\bar\rho) = \begin{cases} \dfrac{e^{\tau(\bar\rho)t/k} - e^{\tau(\bar\rho)(t-1)/k}}{e^{\tau(\bar\rho)} - 1}, & \tau(\bar\rho) \neq 0, \\ \dfrac{1}{k}, & \tau(\bar\rho) = 0, \end{cases} \qquad t = 1, \dots, k. \tag{32}$$

The beta weighting family provides a two-parameter alternative:

$$W_{\theta,\kappa} = \mathrm{Beta}(\theta\kappa, (1-\theta)\kappa), \qquad 0 < \theta < 1, \quad \kappa > 0.$$

Here θ is the mean fractional threshold and κ controls concentration around that threshold. The finite-budget weights are

$$\omega_{t,k}^{(\theta,\kappa)} = F_{\theta\kappa,(1-\theta)\kappa}(t/k) - F_{\theta\kappa,(1-\theta)\kappa}((t-1)/k), \qquad t = 1, \dots, k,$$

where $F_{a,b}$ is the beta distribution function. The moment family is the subfamily $\mathrm{Beta}(\lambda, 1)$, corresponding to $\theta = \lambda/(\lambda+1)$ and $\kappa = \lambda + 1$.

Power means & question-first nonlinear summaries. Weighted power means give a controlled nonlinear extension when a scalar should change the influence of weak or strong regions of the threshold spectrum:

$$M_{i,W,q}(k) := \left(\int_{[0,1]} S_{i,k}(\rho)^q \, dW(\rho) \right)^{1/q}, \qquad q > 0.$$

At finite budget k this reduces to

$$M_{i,\omega,q}(k) = \left(\sum_{t=1}^{k} \omega_{t,k} S_{i,k,t}^q \right)^{1/q}.$$

The case $q = 1$ recovers $U_{i,\omega}(k)$. For fixed nonnegative weights, $M_{i,\omega,q}(k)$ is nondecreasing in q (Hardy et al., 1952). The influence of individual coordinates follows from differentiation: for positive coordinates $x_t = S_{i,k,t}$ and positive weights,

$$\frac{\partial M_{i,\omega,q}/\partial x_t}{\partial M_{i,\omega,q}/\partial x_s} = \frac{\omega_{t,k}}{\omega_{s,k}} \left(\frac{x_t}{x_s} \right)^{q-1}.$$

Thus increasing q increases the relative marginal influence of larger coordinates; decreasing q increases that of smaller coordinates.

The dataset-level nonlinear metric averages the nonlinear score over questions:

$$\overline{M}_{W,q}(k) := \frac{1}{M} \sum_{i=1}^{M} M_{i,W,q}(k), \qquad q > 0,$$

and, for finite weights,

$$\overline{M}_{\omega,q}(k) := \frac{1}{M} \sum_{i=1}^{M} \left(\sum_{t=1}^{k} \omega_{t,k} S_{i,k,t}^q \right)^{1/q}.$$

Posterior uncertainty is obtained by evaluating the full question-first functional on each joint posterior draw:

$$\overline{M}_{\omega,q}^{(h)}(k) = \frac{1}{M} \sum_{i=1}^{M} \left(\sum_{t=1}^{k} \omega_{t,k} \left[S_{i,k,t}\left(p_i^{(h)}\right) \right]^q \right)^{1/q}, \qquad q > 0.$$

For $q \neq 1$, the posterior mean profile and covariance matrix alone are insufficient for these nonlinear posterior summaries; one-dimensional Beta integration or posterior simulation is required.

Profile-level nonlinear compression. A nonlinear scalar can also be formed by first averaging the tail profile over questions and then applying the nonlinear spectrum summary:

$$M_{W,q}^{\mathrm{prof}}(k) := \left(\sum_{t=1}^{k} \omega_{t,k} S_{k,t}^q \right)^{1/q}, \qquad q > 0.$$

This scalar compresses the already averaged dataset-level profile. It should be distinguished from the question-first dataset metric $\overline{M}_{W,q}(k) = M^{-1} \sum_i M_{i,W,q}(k)$. The two coincide when $q = 1$, but for $q \neq 1$ they generally differ because the nonlinear power mean fails to commute with averaging over questions:

$$\frac{1}{M} \sum_{i=1}^{M} \left(\sum_{t=1}^{k} \omega_{t,k} S_{i,k,t}^q \right)^{1/q} \not\equiv \left(\sum_{t=1}^{k} \omega_{t,k} \left(\frac{1}{M} \sum_{i=1}^{M} S_{i,k,t} \right)^q \right)^{1/q}.$$

Posterior uncertainty for $M_{W,q}^{\mathrm{prof}}(k)$ can be estimated by applying the same functional to each joint posterior draw of the dataset-level profile $S_k^{(h)}$.

Quadratic RMS & the L^2 score. When no task-specific utility preference is available, the default diagnostic uses uniform threshold weight in the question-first quadratic RMS spectrum family

$$\mathrm{QRS}_W(k) := \frac{1}{M}\sum_{i=1}^{M}\left(\int_{[0,1]} S_{i,k}(\rho)^2\, dW(\rho)\right)^{1/2}. \tag{33}$$

where W may be changed only when the benchmark pre-specifies a different threshold measure. At finite budget,

$$\mathrm{QRS}_W(k) = \frac{1}{M}\sum_{i=1}^{M}\left(\sum_{t=1}^{k}\omega_{t,k}^{(W)} S_{i,k,t}^2\right)^{1/2}. \tag{34}$$

With uniform W, this is the L^2 spectrum score. For a question i,

$$L_{i,2}(k) := \|S_{i,k}\|_{L^2([0,1])} = \left(\int_0^1 S_{i,k}(\rho)^2\, d\rho\right)^{1/2} = \left(\frac{1}{k}\sum_{t=1}^{k} S_{i,k,t}^2\right)^{1/2},$$

and the dataset-level score is

$$L_2@k := \frac{1}{M}\sum_{i=1}^{M} L_{i,2}(k) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{k}\sum_{t=1}^{k} S_{i,k,t}^2\right)^{1/2}.$$

Thus $L_2@k = \mathrm{QRS}_{\mathrm{Unif}}(k) = \overline{M}_{\omega^{\mathrm{unif}},2}(k)$ with $\omega_{t,k}^{\mathrm{unif}} = 1/k$. Uniform-weight QRS supplies a fixed reference for nonlinear profile summaries. It averages question-level root-mean-square tail probabilities, retaining sensitivity to the distribution of success across questions. For example, at $k = 2$, two questions with success probabilities $(0.5, 0.5)$ have $\mathrm{QRS}_{\mathrm{Unif}}(2) = \sqrt{5}/4$, whereas $(0, 1)$ gives $\mathrm{QRS}_{\mathrm{Unif}}(2) = 0.5$. Both have average accuracy 0.5.

The posterior draw of the uniform L^2 score is

$$L_2^{(h)}@k := \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{k}\sum_{t=1}^{k}\left[S_{i,k,t}\left(p_i^{(h)}\right)\right]^2\right)^{1/2}, \qquad h = 1, \ldots, H.$$

Define

$$\mu_{L_2,k} := \mathbb{E}\left[L_2@k \mid R, R^0, \eta\right], \qquad \sigma_{L_2,k}^2 := \mathrm{Var}\left(L_2@k \mid R, R^0, \eta\right).$$

The Monte Carlo estimates are

$$\widehat{\mu}_{L_2,k} = \frac{1}{H}\sum_{h=1}^{H} L_2^{(h)}@k, \qquad \widehat{\sigma}_{L_2,k}^2 = \frac{1}{H-1}\sum_{h=1}^{H}\left(L_2^{(h)}@k - \widehat{\mu}_{L_2,k}\right)^2.$$

A related m-rollout common-threshold utility is

$$Q_{m,W}(k) := \frac{1}{M}\sum_{i=1}^{M}\int_{[0,1]} S_{i,k}(\rho)^m\, dW(\rho), \qquad m \in \{2, 3, \ldots\}.$$

If $\Theta \sim W$ is an independent random threshold and $X_{i,k}^{(1)}, \ldots, X_{i,k}^{(m)}$ are independent k-sample rollouts on the same question, then

$$\int S_{i,k}(\rho)^m\, dW(\rho) = \Pr\left(X_{i,k}^{(1)} \geq \max\{1, \lceil k\Theta\rceil\}, \ldots, X_{i,k}^{(m)} \geq \max\{1, \lceil k\Theta\rceil\} \mid p_i\right).$$

For $m = 2$, this gives the two-rollout utility

$$Q_{2,W}(k) := \frac{1}{M}\sum_{i=1}^{M}\int_{[0,1]} S_{i,k}(\rho)^2\, dW(\rho). \tag{35}$$

QRS applies the square root before averaging over questions, while $Q_{m,W}$ keeps the replication-probability scale.

Harmonic endpoint balance. For positive discovery and stability endpoints, a weighted harmonic score is

$$H_{i,\alpha}^{\mathrm{end}}(k) = \left(\frac{\alpha}{S_{i,k,1}} + \frac{1-\alpha}{S_{i,k,k}} \right)^{-1}, \qquad \alpha \in [0,1],$$

with value zero when either endpoint with positive weight is zero. The harmonic form is familiar from the F-measure, where it is used to require simultaneous performance on two criteria (van Rijsbergen, 1979). The score should be reported with both endpoints because distinct endpoint pairs can have the same harmonic mean. Its dataset-level version is the question-first average $M^{-1} \sum_i H_{i,\alpha}^{\mathrm{end}}(k)$. Nonlinear functions of already averaged endpoint metrics can obscure question-level heterogeneity.

Reference-profile shortfall. Some evaluations have explicit acceptability requirements, such as a minimum discovery level and a minimum stability level. In that setting, a target-profile shortfall is more interpretable than either an average or a product. Let $b = (b_1, \ldots, b_k)$ be a nonincreasing target profile with $0 \le b_t \le 1$, and let $\lambda_t > 0$ be normalizing or preference weights. Define the question-level shortfall

$$A_{i,b,\lambda}(k) = \max_{1 \le t \le k} \lambda_t \left[b_t - S_{i,k,t} \right]_+ + \epsilon \sum_{t=1}^{k} \lambda_t \left[b_t - S_{i,k,t} \right]_+, \qquad \epsilon > 0.$$

Lower values are better, and $A_{i,b,\lambda}(k) = 0$ exactly when the whole profile meets the target. This construction follows the reference-point and achievement-scalarizing tradition in multiobjective optimization (Miettinen, 1998; Marler & Arora, 2004). For this framework, this shortfall is most suitable for constraint-like reporting, while the tail profile remains the object that shows where any violation occurs. The dataset-level shortfall is $A_{b,\lambda}(k) = M^{-1} \sum_i A_{i,b,\lambda}(k)$, with posterior uncertainty obtained by evaluating this average on each joint posterior draw.

D Simulations: Information and Model Selection

Accuracy and both discovery–stability endpoints can agree while intermediate thresholds favor different models. The full profile retains these distinctions, extending the information-loss analysis in Appendix K.3. In the matched-endpoint simulations, TailPass posterior-mean profiles recover the corresponding model choices from finite response banks and reduce regret relative to rankings fixed across thresholds.

D.1 Information beyond Accuracy and Endpoints

Exhaustive enumeration of nondecreasing row-count vectors $0 \le \gamma_1 \le \cdots \le \gamma_M \le N$ with $\sum_i \gamma_i = MN/2$ gives 526 histograms for $M = N = 8$ and 5,448 for $M = N = 10$. Each histogram is counted once, irrespective of the number of labeled response matrices that realize it. For every even budget $k \le N$, the finite-bank profiles in Equation (6) are compared using integer numerators with common denominator $M\binom{N}{k}$, so equality is exact. Their normalized uniform-threshold area is $1/2$ throughout, by the finite-bank tail-sum identity $k^{-1} \sum_t \widehat{S}_{k,t}^{\mathrm{fs}} = (MN)^{-1} \sum_i \gamma_i$, the counterpart of Equation (10).

At $M = N = k = 8$, the endpoint pair gives only 17 distinct reports, while the full profile distinguishes all 526 histograms (Table 5). For example, the count vectors $(3,3,3,3,3,3,7,7)$ and $(4,4,4,4,4,4,4,4)$ both have accuracy $1/2$, discovery one, and strict stability zero. Their probabilities of at least four correct responses are $1/4$ and one, a gap of 75 percentage points (pp). Adding the strict-majority coordinate to the endpoints increases resolution to 53 reports, still fewer than the full profile.

For $M = N = 8$, the profile distinguishes 525 histograms at $k = 4$ and all 526 at $k = 6 < N$. At $k = 2$, the endpoints constitute the full profile. For $M = N = 10$ at $k = N$, 26 endpoint pairs and 100 endpoint-plus-majority reports correspond to 5,448 distinct full profiles. The largest gap hidden by equal endpoints is 80 pp. At $k < N$, the full profile specifies the subsampled success-count distribution; at $k = N$, it also identifies the bank histogram through Equation (5).

Table 5: Distinct finite-bank reports at accuracy $1/2$: 526 histograms for $M = N = 8$ and 5,448 for $M = N = 10$. The majority control adds the $t = \lfloor k/2 \rfloor + 1$ coordinate to both endpoints. Gap: largest threshold-probability difference between histograms with identical endpoint pairs, in percentage points (pp).

$M = N$	k	Endpoints	Endpoints + majority	Full profile	Gap (pp)
8	2	55	55	55	0.00
8	4	518	525	525	5.36
8	6	374	526	526	37.50
8	8	17	53	526	75.00
10	2	109	109	109	0.00
10	4	4,867	5,347	5,347	9.33
10	6	4,903	5,447	5,448	28.57
10	8	1,277	4,426	5,448	40.00
10	10	26	100	5,448	80.00

D.2 Model Selection with Matched Endpoints

At $k = 16$, let model A's per-question success probability have equal mass at d and $1 - d$. Model B has mass $1 - w$ at $p = 1/2$ and mass $w/2$ at each of $p = 0$ and $p = 1$. Matching strict stability gives

$$w = \frac{\frac{1}{2}\{d^{16} + (1-d)^{16}\} - 2^{-16}}{\frac{1}{2} - 2^{-16}}. \tag{36}$$

Both distributions are symmetric about $1/2$, so discovery also matches and both accuracies are $1/2$ (Equations (7)–(8)). For the primary setting $d = 1/4$, $w \approx 0.009992$. Model B has the higher probability of at least four correct responses, whereas A is better when thirteen are required. The gap is 18.70 pp in each direction (Figure 4).

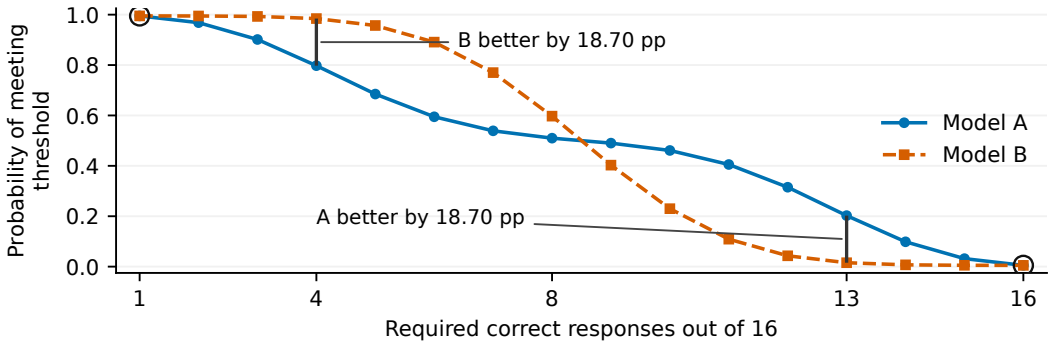

Figure 4: Exact population profiles with equal accuracy and endpoints (circled). The preferred model reverses between requirements of four and thirteen correct responses out of sixteen.

Each of 4,000 independent replications samples $M = 50$ paired questions and $N = 16$ attempts per question and model. For each question, an independent $U_i \sim \mathrm{Unif}(0, 1)$ selects the same quantile of both success-probability distributions: A uses d below $1/2$ and $1 - d$ otherwise; B uses zero below $w/2$, one above $1 - w/2$, and $1/2$ between these cutoffs. Attempts are independent across questions and models conditional on these probabilities.

All selectors receive the same observed banks. At each required threshold, the TailPass selectors choose the model with the larger posterior mean in Equation (4), with independent Beta(α, α) priors for $\alpha \in \{1, 1/2\}$ and no auxiliary evidence. Thus each row posterior is Beta$(\gamma_i + \alpha, N - \gamma_i + \alpha)$; $\alpha = 1$ is the default in Section 3.1, and $\alpha = 1/2$ is a prior sensitivity check.

The finite-bank comparator uses the hypergeometric profile estimator in Equation (6). At $N = k$, each coordinate is the observed fraction of questions meeting threshold t:

$$\widehat{S}^{\mathrm{fs}}_{k,t} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{\gamma_i \geq t\}.$$

Its selector chooses the larger estimate at the required threshold, equivalently evaluating G-Pass@k_τ at $\tau = t/k$ (Liu et al., 2025). Fixed scalar selectors rank by observed accuracy, discovery, strict stability, the uniform discovery-budget average over 1:16, question-first finite-bank QRS, or strict majority (Appendices C.8 and M). Each of these makes one choice for all thresholds. Score differences within 10^{-12} split the choice equally.

The evaluation target averages over fresh questions as well as fresh attempts. Writing F_m for model m's generating distribution, define $s_{m,t} = \mathbb{E}_{p\sim F_m}[q_{16,t}(p)]$ for $m \in \{A, B\}$. The exact population probabilities evaluate the selected model's utility. If $\pi_{b,t}$ is the probability of choosing A in replication b, including a half choice for ties, the regret averaged over a uniform distribution of required thresholds is

$$\mathcal{R}_b = \frac{1}{16}\sum_{t=1}^{16}\left[\max_{m\in\{A,B\}} s_{m,t} - \pi_{b,t}s_{A,t} - (1-\pi_{b,t})s_{B,t}\right]. \tag{37}$$

For any choice fixed across thresholds, the area identity gives mean utility $1/2$, hence regret of 7.449 pp for these populations. TailPass reduces mean regret to 0.690 pp under the uniform prior, a reduction of 90.7% (Table 6). Under both priors, TailPass recovers the opposing choices at $t = 4$ and $t = 13$ in at least 99.95% of replications, counting ties as one half. Monte Carlo standard errors use the 4,000 replication-level values of $\mathcal{R}_b$, preserving dependence across thresholds.

Table 6: Mean regret in percentage points (pp) for $d = 1/4$, $M = 50$, and $N = k = 16$, using the indicated Beta priors for TailPass. Parentheses: Monte Carlo standard errors over 4,000 replications. All six fixed scalar selectors have identical mean regret by the area identity. The profile selectors adapt their choice to the required threshold.

Selector	Regret (pp)
Fixed scalar ranking	7.449
Hypergeometric profile	0.294 (0.005)
TailPass, Beta$(1, 1)$	0.690 (0.014)
TailPass, Beta$(\frac{1}{2}, \frac{1}{2})$	0.692 (0.014)

D.3 Sensitivity Analysis

The full grid uses $d \in \{1/5, 1/4, 3/10, 2/5\}$, $M \in \{50, 200\}$, and $N \in \{16, 32, 64\}$ at $k = 16$, with 4,000 replications per population pair and root seed 2026092507. For each d, question prefixes and nested attempt counts share observations across the size grid. Both TailPass prior settings have lower mean regret than the fixed scalar choices in all 24 population–size settings. At $d = 1/4$ and $M = 50$, increasing N from 16 to 64 lowers the uniform-prior TailPass regret from 0.690 to 0.225 pp.

TailPass uses intermediate-threshold probabilities to recover model choices that accuracy and endpoint reports leave unresolved. The same profile supports joint posterior uncertainty and pre-specified utility summaries (Sections 3.1 and 3.2).

E Choosing Utilities and Reporting Budgets

The utility specifies which outcomes matter; the reporting budget specifies the operation being evaluated. The utility and budget rule are fixed independently of test rankings and shared across compared models. Correctness observations estimate performance under these choices, but cannot determine the value of discovery relative to repeatability.

E.1 Match the utility to the consequence

Let $u_k(x)$ value x correct responses among k attempts, with $u_k(0) = 0$, $u_k(k) = 1$, and u_k nondecreasing. The tail-sum identity gives

$$U_u(k) = \frac{1}{M}\sum_i \mathbb{E}[u_k(X_{i,k})] = \sum_{t=1}^{k}[u_k(t) - u_k(t-1)]S_{k,t}. \tag{38}$$

Thus the threshold weights are the marginal values of additional successes. When one identifiable correct answer suffices, $u_k(x) = \mathbf{1}\{x \geq 1\}$ gives discovery. Equal value per correct execution gives $u_k(x) = x/k$ and accuracy. A requirement of at least r successes gives $S_{k,r}$; all-success reliability gives $S_{k,k}$. A known requirement of 80% correct therefore calls for $S_{k,\lceil .8k \rceil}$. Realizing discovery value requires identifying a correct candidate; an actual verifier or voting policy requires its own evaluation (Appendix L.1).

The moment family $u_k(x) = (x/k)^\lambda$ offers a compact model when value depends smoothly on the correct *fraction*, without a hard acceptance threshold. It assumes constant value elasticity, $zu'(z)/u(z) = \lambda$ for $z > 0$. Values $0 < \lambda < 1$ express diminishing marginal value, $\lambda = 1$ equal marginal value, and $\lambda > 1$ increasing marginal value near complete success. The latter is convex: it rewards variation toward extreme outcomes at fixed mean, so repeatability weighting is not risk aversion and provides no worst-question or worst-field guarantee. Distinct error or resource costs instead require a payoff that includes those outcomes, as in the categorical construction of Section 3.3.

E.2 Elicit the exponent from a value judgment

With other consequences held fixed, assign an interior correct fraction $\rho = t/k \in (0, 1)$ a value $v \in (0, 1)$ relative to zero and complete success. Under expected utility, v can be elicited by indifference between that certain outcome and a lottery yielding complete success with probability v and complete failure otherwise. Then

$$\rho^\lambda = v \quad \Longrightarrow \quad \lambda = \frac{\log v}{\log \rho}. \tag{39}$$

Valuing a half-correct batch at one quarter of a fully correct batch gives $\lambda = 2$; valuing it at one sixteenth gives $\lambda = 4$. A second anchor checks the power assumption. Incompatible anchors call for a richer payoff; imprecise anchors call for a range of exponents. These preferences are separate from posterior uncertainty about performance. Maximizing the score over λ cannot elicit them: z^λ decreases with λ for $0 < z < 1$, mechanically favoring the smallest allowed exponent. When the use case is unspecified, report the profile and a declared sensitivity range, with $\lambda = 1$ as a linear reference.

E.3 Set the budget from the operation and its cost

A production batch or retry allowance supplies k; the observed bank size N determines the evidence available to estimate its performance. Collecting more observations need not change the reporting target. At fixed sample count, compare the same k. Equal sample counts need not imply equal compute; compute-matched evaluation includes generation and verifier costs (Snell et al., 2025). At fixed resources, models may use different k under common accounting that also includes selection costs. Fix and report decoding, stopping, and failure accounting as part of each compared system.

Optimizing a normalized moment score over k does not select a useful resource budget. Under the conditionally independent Bernoulli model,

$$U_2(k) = \frac{1}{M} \sum_i \left[p_i^2 + \frac{p_i(1 - p_i)}{k} \right], \tag{40}$$

so $k = 1$ maximizes this score even when further attempts improve discovery. If deployment permits budget selection, optimize an explicit net payoff. For a fixed batch, value B for one identifiable success and cost c per attempt give $V_m(k) = BS_{k,1}^m - ck$; another attempt is worthwhile when $BM^{-1} \sum_i p_{mi}(1-p_{mi})^k > c$. Heterogeneous costs require expected total expenditure, and stopping on success requires a different cost model. Without a specified operating budget or price, report a common budget curve. Extrapolating the full profile beyond the observed bank requires a sampling model and corresponding assumption and prior checks.

E.4 Fix reporting choices before comparing models

The population weights, utility anchors, primary budget or resource policy, and sensitivity range define a common comparison. Select models or budget policies on development data and evaluate the

frozen choice on held-out data. Optimizing a noisy selection criterion can bias the reported result. Inferential claims selected from a grid require a separate final evaluation or simultaneous uncertainty across the comparisons; an interval for the selected cell alone does not account for selection. Held-out evaluation addresses statistical optimism, but does not establish that the selected utility reflects a practitioner's preferences.

The SuperGPQA banks in Appendix H contain 3,600 questions in 72 fields, with 80 attempts per question and configuration. At $k = 16$, varying $\lambda \in \{.25, .5, 1, 2, 4, 8\}$ changes the estimated leader in 2/72 fields among all four configurations, and 38/72 among the three GPT-OSS effort configurations. These are descriptive sensitivities, conditional on the candidate set. For the high-minus-medium contrast, selecting each field's most favorable cell over these exponents and $k \in \{1, 2, 4, 8, 16, 32\}$ gives a mean fitting advantage of 6.39 percentage points and a held-out advantage of 5.00. Each of 128 partitions separates 25 fitting from 25 evaluation questions per field and 40 fitting from 40 evaluation attempt positions; utilities average all size-k subsets exactly. The 1.39-point optimism has a 95% paired field-bootstrap interval of [1.19, 1.59], after averaging partitions within each field. A fixed $(k = 16, \lambda = 2)$ gives fitting and held-out differences of 4.29 and 4.34 points. Reused partitions are not independent replications, and these response banks contain no stakeholder preference judgments from which to infer a universally appropriate exponent or budget.

F Extended Related Work

Repeated sampling & inference-time compute. Code-generation benchmarks established repeated sampling as an evaluation protocol for test-based functional correctness (Chen et al., 2021). AlphaCode extended this protocol to large candidate pools with filtering, clustering, and ranking before submission (Li et al., 2022), and EvalPlus showed that incomplete tests can accept programs that fail stronger test suites (Liu et al., 2023a). Reasoning systems aggregate sampled answers by self-consistency (Wang et al., 2023), select complete solutions with learned verifiers (Cobbe et al., 2021), train verifiers with intermediate-step feedback (Lightman et al., 2024), search over reasoning states (Yao et al., 2023), and scale inference computation (Brown et al., 2024; Snell et al., 2025). Surveys organize these inference regimes and their evaluation (Zhang et al., 2025; Hariri et al., 2026a). Across these settings, the evaluator observes a bank of candidate outputs, and the reported metric determines which event in that bank receives credit.

Discovery, voting, & stability. Discovery-oriented scores match settings in which a trusted evaluator can inspect a batch and retain an accepted candidate. Self-consistency selects an answer by aggregating answers extracted from sampled reasoning paths (Wang et al., 2023). Agent benchmarks such as τ-bench measure repeated-execution reliability in tool-use environments (Yao et al., 2025). G-$\mathrm{Pass@}k_\tau$ makes the required reliability level explicit (Liu et al., 2025). $\mathrm{TailPass@}k$ reports $\Pr(X_k \geq t)$ for every threshold t, so $\mathrm{pass@}k$ $(t = 1)$, majority-style thresholds, fixed-fraction reliability, and all-k repeatability are read from one success-count tail rather than selected as separate endpoints.

Uncertainty & latent success profiles. Finite response banks leave sampling uncertainty in both point estimates and rankings. The combinatorial estimator averages dependent subsets of the same bank (Chen et al., 2021); Bayesian LLM-evaluation methods quantify uncertainty in binary outcomes (Longjohn et al., 2025) and in rubric-weighted scores and rankings (Hariri et al., 2026c;b). For an item with latent success probability p_i, discovery applies $1 - (1 - p_i)^k$, while majority voting applies a binomial upper-tail polynomial. For binary correctness under conditionally independent, identically distributed calls, mean one-call accuracy does not identify the gains available from majority voting (Liu, 2026). $\mathrm{TailPass@}k$ applies the same tail-probability view across all thresholds and carries posterior uncertainty for the coordinates jointly, which matters when model rankings depend on the chosen threshold.

Labels, verifiers, & utilities. Repeated-sampling metrics inherit the measurement error of their labels. Incomplete tests can accept programs that fail stronger tests (Liu et al., 2023a); resampling with imperfect verifiers can increase the risk of returning a false positive (Stroebl et al., 2024); learned verifiers score complete solutions (Cobbe et al., 2021), while process supervision supplies feedback on intermediate steps (Lightman et al., 2024). Zheng et al. (2023) find position and verbosity biases

in LLM judges, but their evidence for self-enhancement bias is inconclusive. These errors enter through the labeled response record, before any threshold profile or scalar summary is computed. Joint evaluation of discovery, repeatability, cost, latency, and harm requires explicit trade-offs, as in multiattribute decision models (Keeney & Raiffa, 1976). G-Pass@k_τ fixes one reliability threshold (Liu et al., 2025); Bayesian rubric evaluation places uncertainty on weighted category probabilities (Hariri et al., 2026c); and binary vote-accuracy curves characterize majority-vote gains under conditionally independent, identically distributed calls (Liu, 2026). TailPass@k uses the threshold tail itself as the estimand and scalarizes only after estimating joint uncertainty over that profile.

G Extended Experiment Analyses

G.1 Convergence & Rank Correlation

Convergence is measured against the model ordering of the full response bank. With equal trial counts, full-budget average accuracy is order-equivalent to the uniform-prior Bayesian mean used as a reference in prior evaluation work (Hariri et al., 2026c). For each sampled set of models, the reference is the competition ranking induced by Avg@80 on that same sampled set, as in the model-pool comparisons of Hariri et al. (2026b). The comparison therefore asks how quickly a metric applied to prefixes of the response bank recovers the full-budget average-accuracy ordering. This target corresponds to uniform weighting over correctness thresholds by the tail-sum identity.

G.1.1 Matched Resampling

The response-bank tensors and sampling protocol are specified in Section J.2. The same full-budget reference definition is used in both experiments.

For each task, model-subset size $s_{\mathrm{mod}} \in \{5, 10, 15\}$, and sampling regime, 1000 model-set replicates are drawn. Bootstrap replicates sample models with replacement (Efron, 1979), while subsampling replicates sample without replacement. Duplicate model identities in bootstrap replicates are retained and can induce ties. For each task, model-subset size, and regime, a shared bank of 1000 random trial-order permutations is generated from seed 42 and reused by every metric. Thus every method is evaluated on the same 10^6 matched subset-permutation realizations per configuration, with $4 \times 3 \times 2 = 24$ configurations in total.

The metric grid includes average accuracy, pass@k, Unanimous@k, AUC@k, Maj@k, generalized threshold-pass baselines, and TailPass summaries at $k \in \{2, 4, 8\}$.

Prefix scores. For prefix n, each model–question row is restricted to the first n trials under the shared trial-order permutation, and $\gamma_i(n)$ denotes the corresponding number of correct responses for question i. The prefix-average row is denoted by Avg@n. For a fixed reporting budget $k \le n$, the finite-bank pass@k, Unanimous@k, G-Pass@k_τ, Maj@k, and AUC@k baselines are computed from the hypergeometric finite-sample profile in Equation (6), with the observed-run count set to n and γ_i replaced by $\gamma_i(n)$. pass@k is the $t = 1$ coordinate, Unanimous@k is the $t = k$ coordinate, G-Pass@k_τ uses $t = \lceil \tau k \rceil$, and Maj@k uses $t = \lfloor k/2 \rfloor + 1$. Throughout this subsection, including its tables and figures, AUC@k denotes the normalized trapezoidal discovery-budget average

$$\mathrm{AUC@}k = \begin{cases} \mathrm{pass@}1, & k = 1, \\ \dfrac{\frac{1}{2}\mathrm{pass@}1 + \sum_{j=2}^{k-1} \mathrm{pass@}j + \frac{1}{2}\mathrm{pass@}k}{k-1}, & 2 \le k \le n, \end{cases} \tag{41}$$

where every pass@j uses the same n-trial prefix and the sum is empty when $k = 2$. For $k > 1$, the endpoint weights are $1/[2(k-1)]$, the interior weights are $1/(k-1)$, and the weights sum to one. The metric analysis and response-bank diagnostics instead use the uniform discrete average $k^{-1} \sum_{j=1}^{k} \mathrm{pass@}j$ (Appendix M); the two weighting schemes coincide for $k = 1, 2$. The finite-bank discovery estimator is used in code-generation evaluation (Chen et al., 2021), all-success scores in agent reliability evaluation (Yao et al., 2025), and intermediate-threshold scores in reasoning stability evaluation (Liu et al., 2025).

TailPass-Moment applies the Beta–Bernoulli posterior from Section 3.1 to the same prefix counts under the uniform $\mathrm{Beta}(1,1)$ prior, then uses the moment-family weights in Equation (29). Small λ emphasizes discovery thresholds, large λ emphasizes stability thresholds, and $\lambda = 1$ gives uniform threshold weight, as in Appendix C.8. Under the uniform $\mathrm{Beta}(1,1)$ prior, the $\lambda = 1$ posterior mean is an affine transformation of $\mathrm{Avg@}n$, so the TailPass-Moment rows with $\lambda = 1$ are rank-equivalent to prefix average accuracy at every valid prefix. The three k values differ only in their minimum valid prefix for convergence; on the common Kendall window $\{8, \ldots, 80\}$, they induce identical rankings.

TailPass-QRS uses the question-first quadratic RMS spectrum summary in Equations (33)–(34), evaluated on the same prefix posterior. It is included as a secondary nonlinear spectrum diagnostic because the square root is applied at the question level before averaging over questions.

G.1.2 Convergence

For a fixed task, sampled model set, trial-order permutation, and scalar metric g, let $r_n^{(g)}$ be the competition ranking after revealing n trials, and let $r^\star$ be the $\mathrm{Avg@80}$ reference ranking on the same sampled model set. The convergence trial is

$$C_g = \min\{n \in \{1, \ldots, 80\} : r_{n'}^{(g)} = r^\star \text{ for all } n' = n, \ldots, 80\}.$$

If the set is empty, the realization is assigned the sentinel value 81. For a k-parameterized metric, prefixes $n < k$ are unavailable, so convergence begins no earlier than trial k. The empirical mass function over the 10^6 matched realizations in one configuration is

$$p_g(n) = \Pr(C_g = n), \qquad n = 1, \ldots, 80, \qquad p_{\mathrm{nc}} = \Pr(C_g = 81).$$

The reported summaries are the penalized mean convergence trial

$$\mu_{\mathrm{pen}} = \sum_{n=1}^{80} n\, p_g(n) + 81 p_{\mathrm{nc}},$$

and the finite mean conditional on convergence,

$$\mu_{\mathrm{fin}} = \frac{\sum_{n=1}^{80} n\, p_g(n)}{1 - p_{\mathrm{nc}}}, \qquad p_{\mathrm{nc}} < 1.$$

When $p_{\mathrm{nc}} = 1$, no sampled realization converges: the conditioning event has zero empirical probability, so μ_{fin} is undefined. Aggregate summaries average each quantity separately with equal weight across the 24 configurations; the finite-mean average includes only configurations with $p_{\mathrm{nc}} < 1$. Configurations with no convergence remain in the μ_{pen} and p_{nc} averages, contributing 81 and 1, respectively. The identity $\mu_{\mathrm{pen}} = (1 - p_{\mathrm{nc}})\mu_{\mathrm{fin}} + 81 p_{\mathrm{nc}}$ holds within each configuration with finite convergence but need not hold between the separately averaged columns. Methods are ordered by lower μ_{pen}, then lower p_{nc}, then lower μ_{fin}.

Table 7 shows that prefix average accuracy has the lowest aggregate penalized mean, as expected from the full-budget average-accuracy target. Among non-reference rows, the uniform-threshold TailPass moment utilities are closest to the reference and have zero non-convergence mass across the aggregate. The strongest non-TailPass baseline in this summary is $\mathrm{AUC@2}$, with $\mu_{\mathrm{pen}} = 54.00$ and $p_{\mathrm{nc}} = 0.461$. The zero non-convergence mass of TailPass-Moment with $\lambda = 1$ follows from its rank equivalence to prefix average accuracy.

G.1.3 Kendall's τ_b Rank Correlation

The Kendall experiment uses the same model subsets, trial-order permutations, and $\mathrm{Avg@80}$ reference rankings. For each prefix ranking and reference ranking, Kendall's τ_b is computed with tie correction (Kendall, 1945). For one matched realization e, let N_{con} and N_{dis} be the numbers of concordant and discordant unordered model pairs, and let N_x and N_y be the numbers of pairs tied only in the prefix ranking or only in the reference ranking. The prefix statistic is

$$\kappa_{b,e,n}^{(g)} = \frac{N_{\mathrm{con}} - N_{\mathrm{dis}}}{\sqrt{(N_{\mathrm{con}} + N_{\mathrm{dis}} + N_x)(N_{\mathrm{con}} + N_{\mathrm{dis}} + N_y)}}.$$

If the denominator is zero, the prefix value is left undefined. The mean prefix-wise curve $\bar{\kappa}_{b,n}^{(g)}$ averages only defined values over the matched subset-permutation realizations.

Table 7: Aggregate convergence summary over all 24 configurations. For each realization, convergence is the first prefix after which the metric-induced ranking remains equal to the Avg@80 reference ranking through trial 80; non-converged realizations are assigned sentinel value 81. μ_{pen} averages this sentinel-penalized convergence trial, p_{nc} is the non-convergence mass, and μ_{fin} conditions on finite convergence. Each column is an unweighted mean of its configuration-level values. Only the μ_{fin} average omits configurations with no converged realizations, for which the conditional mean is undefined (one for pass@2, four for pass@8); all 24 configurations contribute to μ_{pen} and p_{nc}. Lower μ_{pen} and lower p_{nc} are better.

Method	μ_{pen}	p_{nc}	μ_{fin}
Avg@n	47.74	0.000	47.74
$U_{\lambda=1}(2)$	47.79	0.000	47.79
$U_{\lambda=1}(4)$	47.98	0.000	47.98
$U_{\lambda=1}(8)$	48.57	0.000	48.57
L_2@2	48.99	0.296	40.56
AUC@2	54.00	0.461	38.97
Maj@2	57.59	0.466	44.13
G-Pass@$4_{0.5}$	59.79	0.601	38.58
pass@2	60.67	0.621	38.95
pass@8	71.91	0.827	41.03

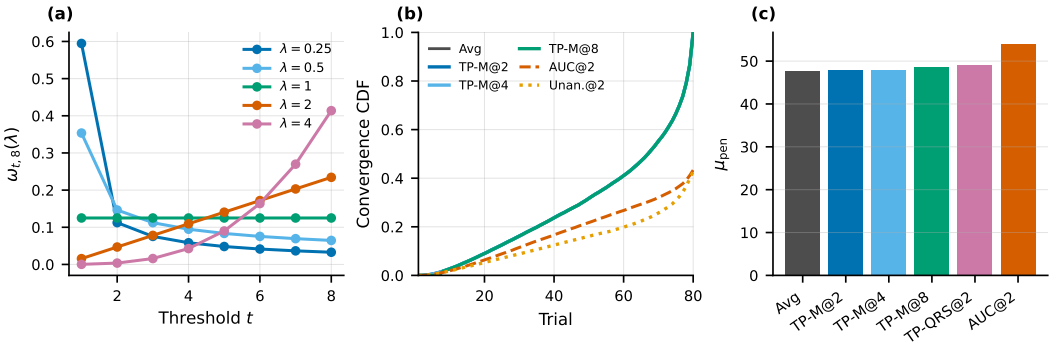

Figure 5: TailPass convergence diagnostics. (a) Moment-family threshold weights for $k = 8$. (b) Matched convergence CDFs under subsampling with subset size 10; TP-M@k denotes TailPass-Moment@k with $\lambda = 1$. (c) Aggregate penalized mean convergence trial over all 24 configurations. Lower μ_{pen} is better, and higher CDF values indicate earlier convergence. The TP-QRS@2 label denotes the spectrum score L_2@2.

All retained metrics are defined by prefix 8, so the common comparison window is $\mathcal{P}_{\text{common}} = \{8, \ldots, 80\}$. The primary summary is the normalized trapezoidal area

$$\text{AUC}^{(g)}_{\tau_b,8:80} = \frac{\frac{1}{2}\bar{\kappa}^{(g)}_{b,8} + \sum_{n=9}^{79} \bar{\kappa}^{(g)}_{b,n} + \frac{1}{2}\bar{\kappa}^{(g)}_{b,80}}{80 - 8}.$$

The endpoint summary is $\bar{\kappa}^{(g)}_{b,80}$, and the undefined mass at a prefix is the fraction of matched realizations with undefined τ_b.

The rank-correlation results agree with the convergence experiment. On the common prefix window, the uniform-threshold moment utilities $U_{\lambda=1}(k)$ match Avg@n, giving $\text{AUC}_{\tau_b,8:80} = 0.9657$ and exact agreement with the Avg@80 reference at prefix 80. The strongest non-TailPass baseline in the aggregate τ_b summary is AUC@2 ($\text{AUC}_{\tau_b,8:80} = 0.9519$). Higher-budget discovery endpoints such as pass@8 are substantially less aligned with the full-budget average-accuracy reference.

G.2 TailPass@k Uncertainty

Appendix J.4 gives the response-bank shape, posterior, draw count, and ranking rule used for Section 4.2. The primary scalar view is the moment-family utility $U_\lambda(8)$ with $\lambda = 2$, which gives

Table 8: Moment-utility sensitivity in the convergence experiment. All rows use $U_\lambda(k)$; lower μ_{pen} is better.

k	λ	μ_{pen}	p_{nc}
2	0.25	55.41	0.509
2	0.50	52.65	0.421
2	1.00	47.79	0.000
2	2.00	51.62	0.385
2	4.00	55.00	0.447
4	0.25	61.15	0.625
4	0.50	54.70	0.510
4	1.00	47.98	0.000
4	2.00	53.95	0.435
4	4.00	61.28	0.644
8	0.25	65.48	0.737
8	0.50	56.84	0.558
8	1.00	48.57	0.000
8	2.00	55.78	0.447
8	4.00	63.27	0.665

Table 9: Aggregate Kendall's τ_b summary over all 24 configurations. Higher $\mathrm{AUC}_{\tau_b,8:80}$ and higher $\bar{\kappa}_{b,80}$ are better.

Method	$\mathrm{AUC}_{\tau_b,8:80}$	$\bar{\kappa}_{b,80}$	$p_{\mathrm{undef},80}$
Avg@n	0.9657	1.0000	0.000
$U_{\lambda=1}(2)$	0.9657	1.0000	0.000
$U_{\lambda=1}(4)$	0.9657	1.0000	0.000
$U_{\lambda=1}(8)$	0.9657	1.0000	0.000
AUC@2	0.9519	0.9655	0.000
Maj@4	0.9279	0.9404	0.000
G-Pass@$4_{0.5}$	0.9279	0.9392	0.000
pass@2	0.9234	0.9361	0.000
pass@8	0.7953	0.8182	0.000

Table 10: Moment-utility sensitivity under Kendall's τ_b. All rows use $U_\lambda(k)$ with the displayed budget k and moment exponent λ. Higher $\mathrm{AUC}_{\tau_b,8:80}$ and higher $\bar{\kappa}_{b,80}$ are better.

k	λ	$\mathrm{AUC}_{\tau_b,8:80}$	$\bar{\kappa}_{b,80}$
2	0.25	0.9471	0.9603
2	0.50	0.9567	0.9708
2	1.00	0.9657	1.0000
2	2.00	0.9510	0.9662
2	4.00	0.9374	0.9529
4	0.25	0.9224	0.9339
4	0.50	0.9485	0.9604
4	1.00	0.9657	1.0000
4	2.00	0.9421	0.9555
4	4.00	0.9150	0.9169
8	0.25	0.9011	0.8973
8	0.50	0.9425	0.9499
8	1.00	0.9657	1.0000
8	2.00	0.9374	0.9529
8	4.00	0.9027	0.9067

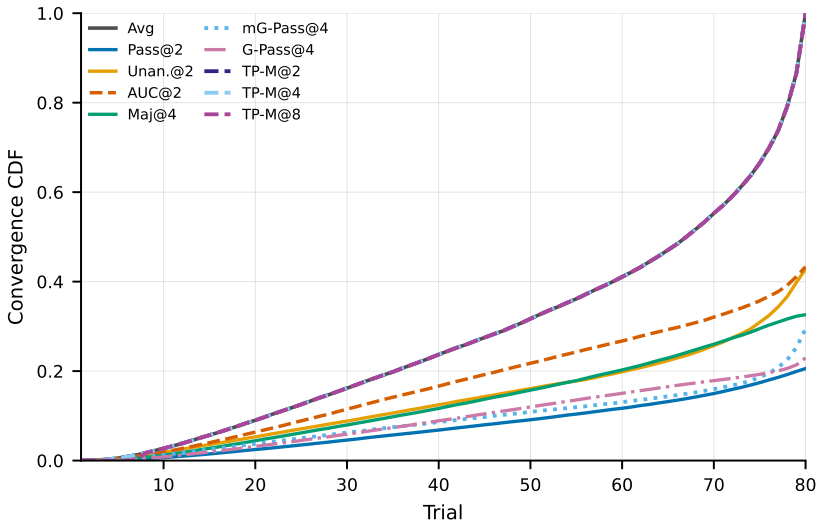

Figure 6: Convergence CDF under subsampling with subset size 10, aggregated over the four tasks using matched model subsets and trial-order permutations. TP-M@k denotes TailPass-Moment@k with $\lambda = 1$, and the G-Pass@$4_{0.5}$ curve uses threshold parameter $\tau = 0.5$.

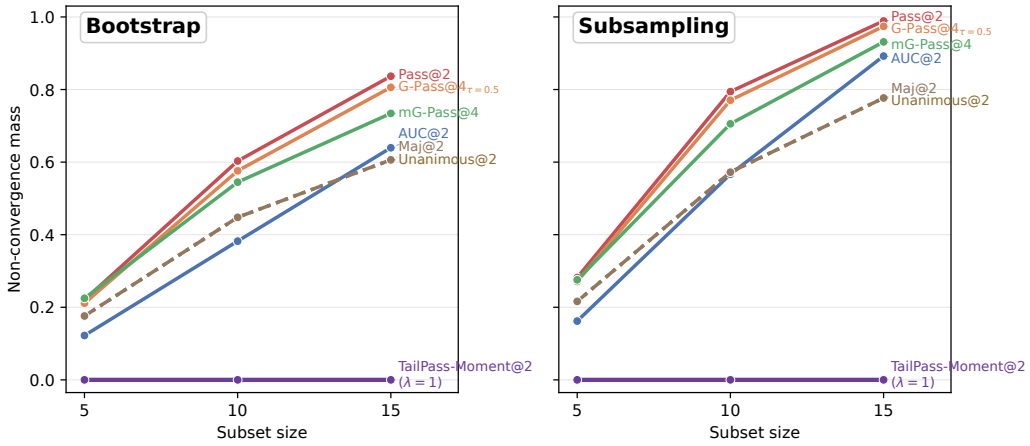

Figure 7: Task-aggregated non-convergence mass for one representative non-reference metric from each retained family. The left panel is bootstrap and the right panel is subsampling. The TailPass-Moment@2 denotes $U_{\lambda=1}(2)$. Lower values are better.

increasing marginal weight to higher correctness thresholds while averaging over the full TailPass@8 profile. For each task, the strict ranking orders models by the posterior mean computed from all 80 top-p trials under this scalar view. The uncertainty-aware ranking keeps the same posterior but groups adjacent models when the 95% equal-tailed credible interval for their pairwise scalar difference contains zero. The all-pairs unresolved count is computed before adjacent grouping; the displayed rank groups are the transitive closure of unresolved adjacent comparisons after sorting by posterior mean.

Table 11 gives the aggregate all-task ranking shown in Figure 10. The strict rank column is a point-estimate order. The UQ rank column applies the pairwise credible-interval rule from Section 4.2; for example, the four models occupying strict ranks 7 through 10 share UQ rank 6. The displayed intervals are marginal scalar intervals, while the unresolved pair decisions are made from posterior draws of pairwise differences. This shared block spans posterior means from 0.518 for OpenThinker2 to 0.504 for Phi-4. The strict order therefore ranks models within a range where the adjacent pairwise posterior intervals do not exclude zero at the chosen credibility level.

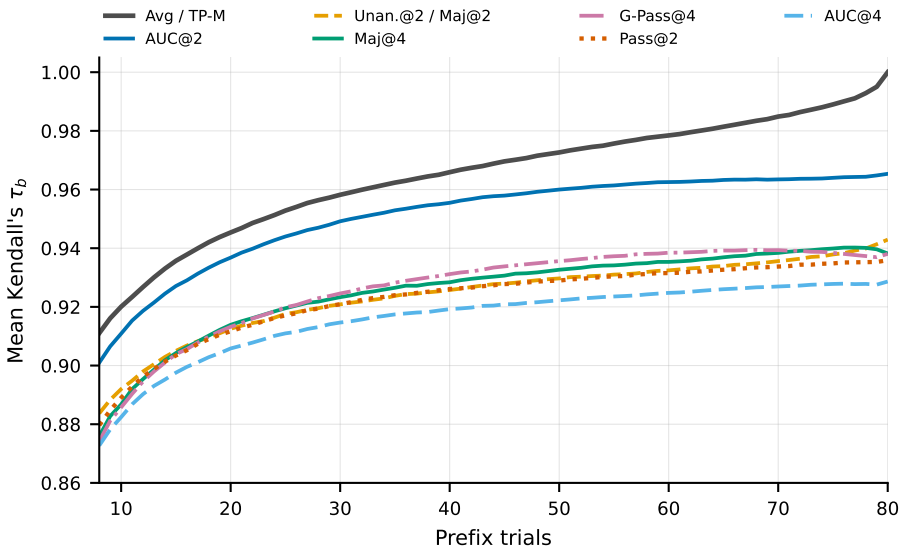

Figure 8: Mean prefix-wise Kendall's τ_b curves under subsampling with subset size 10. Curves use the same sampled model sets and trial-order permutations for every metric. The y-axis is restricted to the observed high-agreement range. Avg / TP-M denotes the coincident $\mathrm{Avg@}n$ and $U_{\lambda=1}(k)$ curves for $k \in \{2, 4, 8\}$; $\mathrm{Unan@2}$ / $\mathrm{Maj@2}$ denotes the coincident $\mathrm{Unanimous@2}$ and $\mathrm{Maj@2}$ curves.

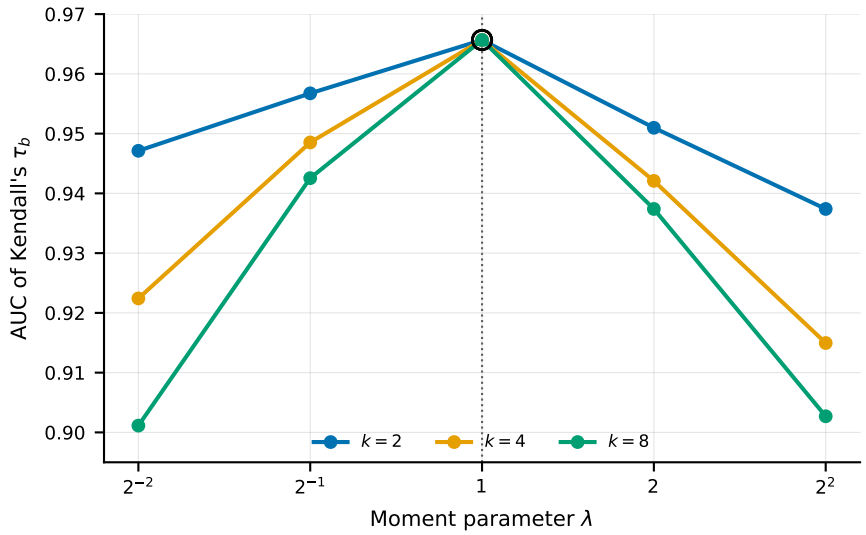

Figure 9: Moment-utility sensitivity of $U_\lambda(k)$ for Kendall's τ_b. The uniform member $\lambda = 1$ is the best member of the scanned grid for $k = 2, 4, 8$ under the full-budget $\mathrm{Avg@80}$ reference. The vertical reference line and open marker indicate $\lambda = 1$.

The profile-width summaries separate discovery uncertainty from stability uncertainty. Aggregate and three of the four individual tasks have their widest median coordinate interval at the stability endpoint $t = 8$, whereas HMMT'25 has its largest median width at the discovery endpoint $t = 1$. A single scalar interval would collapse these endpoint-specific differences. In the aggregate bank, the median interval width is 0.019 at discovery, 0.016 at the middle coordinate $t = 4$, and 0.024 at strict stability. The task-specific profiles show larger posterior scale because each task has only $M = 30$ questions; the largest median widths are 0.057 for AIME'24 and 0.054 for BrUMO'25 at $t = 8$, and 0.051 for HMMT'25 at $t = 1$.

Table 13 holds the response bank, posterior prior, reporting budget, and posterior draws fixed across values of λ. The top aggregate pair is separated for every tested utility, $\lambda \in \{0.25, 0.5, 1, 2, 4, 8\}$.

Table 11: Aggregate all-task ranking by utility $U_{\lambda=2}(8)$ under the no-prior posterior. UQ ranks apply the interval-aware adjacent tie rule from Table 2. Model short names are expanded in Table 35.

Model	Mean	95% CI	Strict rank	UQ rank
Qwen3-Thinking	0.701	[0.695, 0.707]	1	1
Qwen3-4B	0.603	[0.597, 0.609]	2	2
Phi-4-plus	0.570	[0.563, 0.577]	3	3
gpt-oss-high	0.549	[0.541, 0.556]	4	4
AR-Nemotron	0.538	[0.531, 0.545]	5	5
gpt-oss-medium	0.535	[0.527, 0.543]	6	5
OpenThinker2	0.518	[0.511, 0.526]	7	6
FuseO1-DS-QwQ-SkyT1	0.510	[0.502, 0.517]	8	6
Light-R1-DS	0.505	[0.498, 0.512]	9	6
Phi-4	0.504	[0.497, 0.512]	10	6
NVIDIA-Nemotron	0.468	[0.461, 0.475]	11	7
LIMO-v2	0.452	[0.445, 0.459]	12	8
gpt-oss-low	0.376	[0.369, 0.384]	13	9
OR1-Distill	0.357	[0.351, 0.364]	14	10
OpenThinker3	0.357	[0.350, 0.363]	15	10
EXAONE-4.0	0.356	[0.349, 0.363]	16	10
OR-Nemotron	0.325	[0.318, 0.332]	17	11
Sky-T1-Flash	0.195	[0.189, 0.200]	18	12
DS-R1-Qwen	0.183	[0.177, 0.189]	19	13
Bespoke	0.130	[0.125, 0.135]	20	14

Table 12: Median 95% credible-interval width across models for $\mathrm{TailPass@8}$ profile coordinates $S_{8,t}$. The middle column reports $t = 4$.

Task	Discovery width	Middle width	Stability width	Max-width threshold
All	0.019	0.016	0.024	$t = 8$
AIME'24	0.030	0.029	0.057	$t = 8$
AIME'25	0.037	0.033	0.045	$t = 8$
HMMT'25	0.051	0.025	0.034	$t = 1$
BrUMO'25	0.034	0.031	0.054	$t = 8$

Table 13: Viewpoint-dependent uncertainty for aggregate moment utilities $U_\lambda(8)$. Small λ emphasizes discovery thresholds, $\lambda = 1$ is uniform threshold area, and larger λ emphasizes stability-oriented thresholds. Model short names are expanded in Table 35.

λ	Median CI width	Unresolved pairs	Rank groups	τ_b	Top UQ group
0.25	0.014	6/190	14	0.981	Qwen3-Thinking
0.5	0.012	7/190	14	0.976	Qwen3-Thinking
1	0.012	7/190	13	0.976	Qwen3-Thinking
2	0.014	8/190	14	0.973	Qwen3-Thinking
4	0.017	11/190	12	0.954	Qwen3-Thinking
8	0.020	13/190	11	0.940	Qwen3-Thinking

The unresolved-pair count increases from 6 at $\lambda = 0.25$ to 13 at $\lambda = 8$, as both the interval widths and the separation between model scores vary with the threshold weights. The median scalar interval width is smallest for $\lambda \in \{0.5, 1\}$, both 0.012, and increases to 0.020 at $\lambda = 8$. Since every row in the table uses the same response bank and posterior draws, these changes come from the declared threshold weights applied to a common uncertain profile.

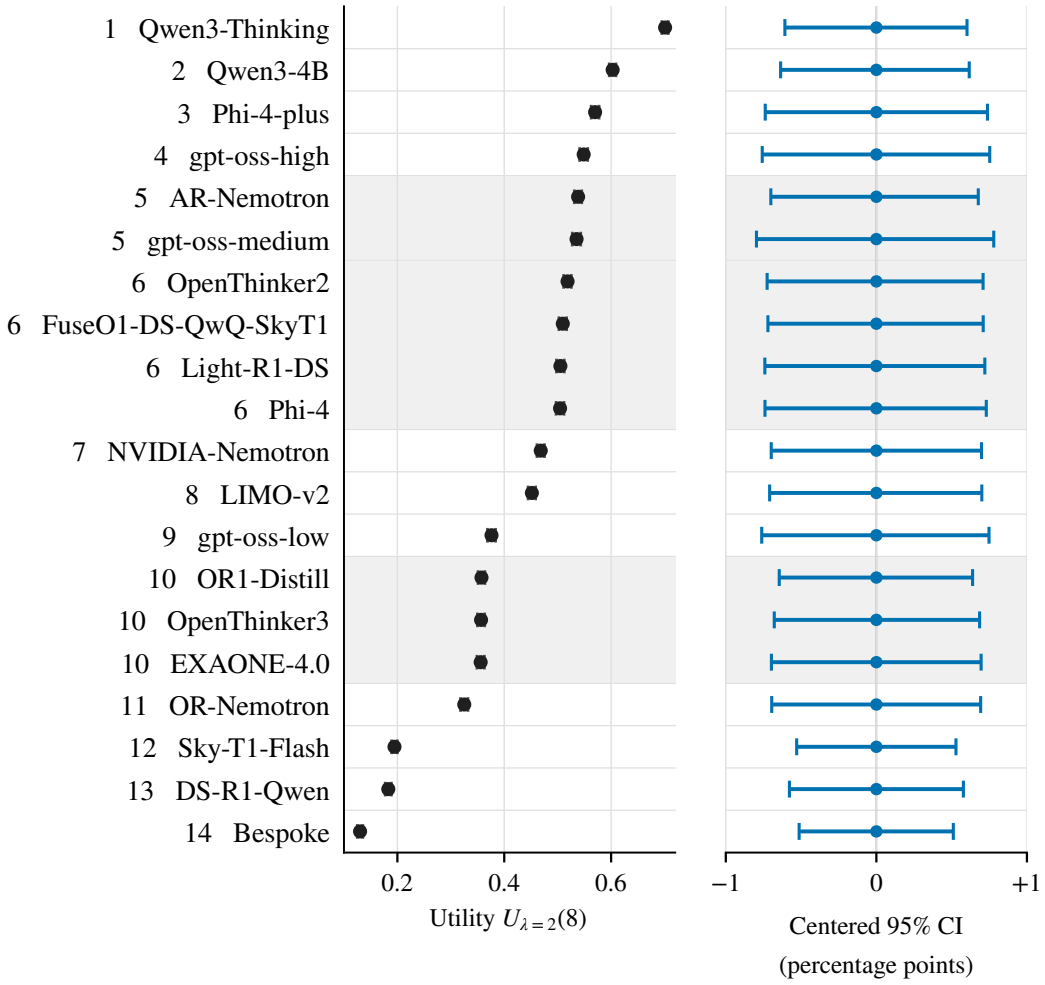

Figure 10: Aggregate interval-aware ranking by utility $U_{\lambda=2}(8)$. Left: posterior means and marginal 95% credible intervals on the utility scale. Right: the same intervals after subtracting each model's posterior mean and multiplying by 100, so one unit is one percentage point of utility. Models are ordered by posterior mean. The number before each model name is the uncertainty-aware dense rank; repeated numbers and shaded blocks indicate groups formed by unresolved adjacent comparisons under the pairwise 95% credible-interval rule. The posterior supports a coarse aggregate ordering: the top models and lower tail are separated, while strict ranks 7–10 collapse into one uncertainty-aware block.

G.3 Prior-Evidence

G.3.1 Validation, Alignment, and Full-bank Sensitivity

Aggregate validation. Appendix J.3 specifies the paired held-out validation protocol. The aggregate bank fixes the reporting budget at $k = 16$ unless stated otherwise, and $n \leq 32$ leaves at least 48 held-out top-p trials per question for the finite-bank reference.

Table 14 gives the aggregate sensitivity grid. At $n = 1$ and $n = 4$, increasing η monotonically lowers held-out MAE for all three moment-family utilities. At $n = 16$ and $n = 32$, the discovery-weighted summary ($\lambda = .5$) still benefits from the greedy prior, but the uniform and repeatability-weighted summaries are best with no transfer. The greedy prior acts as low-sample evidence whose value declines as direct top-p evidence accumulates.

The intervals in Table 3 are narrower with greedy transfer in the reported conditions. The held-out containment rates are well below 95%. These rates measure how often an interval for the latent utility

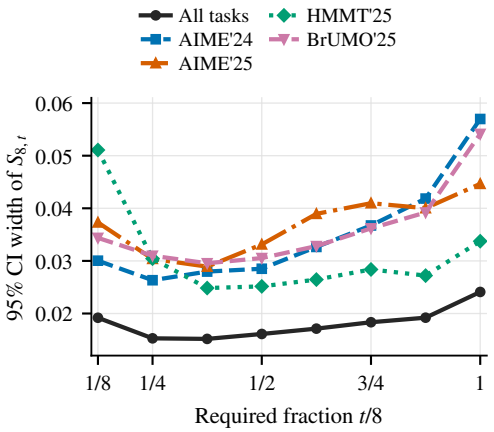

Figure 11: Median 95% credible-interval width of each profile coordinate $S_{8,t}$ across 20 models. Curves compare the four tasks and the 120-question aggregate. The threshold $t/8$ ranges from discovery to strict stability.

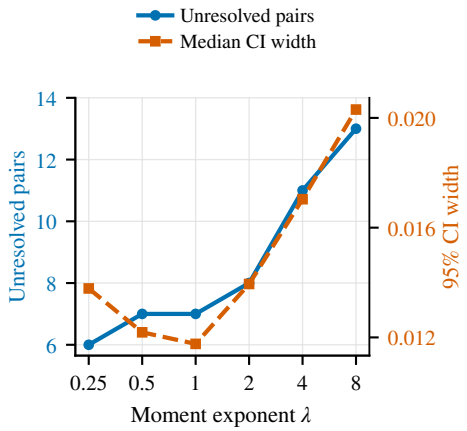

Figure 12: Uncertainty in aggregate utilities $U_\lambda(8)$. Solid blue: unresolved pairs out of 190, whose 95% interval for the utility difference contains zero (left axis). Dashed orange: median marginal 95% interval width across models (right axis). The response bank, posterior draws, and budget are fixed.

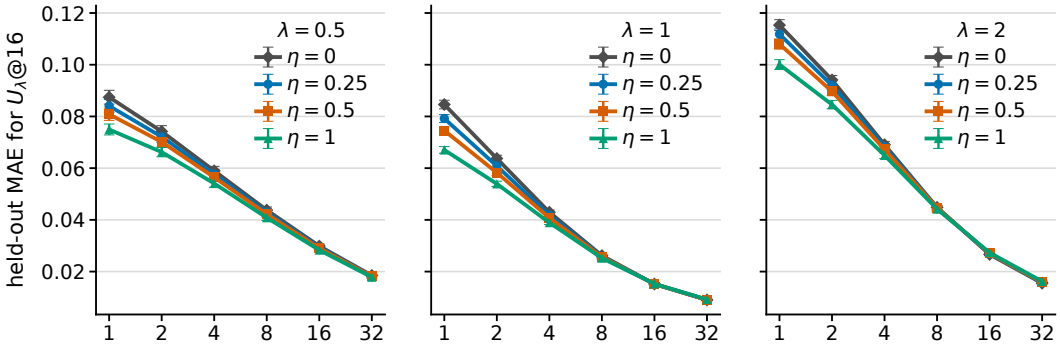

Figure 13: Held-out mean absolute error for aggregate U_λ@16 summaries as a function of the number of observed top-p trials per question. The separation among transfer weights is largest in the small-n regime, and the curves merge as the stochastic response bank becomes more informative.

Table 14: Aggregate held-out validation over prior weights. Entries are mean absolute error to the held-out top-p finite-bank reference for U_λ@16.

n	λ	$\eta = 0$	$\eta = .25$	$\eta = .5$	$\eta = 1$	best η
1	0.5	0.0874	0.0841	0.0808	0.0750	1
1	1	0.0846	0.0792	0.0746	0.0670	1
1	2	0.1153	0.1118	0.1080	0.1001	1
4	0.5	0.0590	0.0577	0.0564	0.0541	1
4	1	0.0429	0.0418	0.0408	0.0389	1
4	2	0.0691	0.0682	0.0672	0.0650	1
16	0.5	0.0299	0.0295	0.0291	0.0283	1
16	1	0.0152	0.0152	0.0152	0.0153	0
16	2	0.0266	0.0268	0.0270	0.0273	0
32	0.5	0.0185	0.0183	0.0182	0.0178	1
32	1	0.0090	0.0091	0.0091	0.0092	0
32	2	0.0155	0.0157	0.0159	0.0162	0

contains the utility of a realized finite bank. Their interpretation depends on the sampling variability of that reference, the prior, and the normal approximation used for the intervals. At very small n, narrower intervals coincide with lower held-out MAE. Once $n \geq 16$, the same narrowing does not improve point prediction for the uniform or repeatability-weighted summaries.

Table 15: Per-task effect of full greedy-prior transfer. Entries are ΔMAE relative to $\eta = 0$ for U_λ@16; negative values indicate that the greedy prior improves held-out top-p prediction.

Task	$n = 4$			$n = 16$		
	$\lambda = .5$	$\lambda = 1$	$\lambda = 2$	$\lambda = .5$	$\lambda = 1$	$\lambda = 2$
AIME'24	-0.0059	-0.0034	-0.0038	-0.0014	0.0006	0.0015
AIME'25	-0.0045	-0.0029	-0.0033	-0.0012	0.0003	0.0010
HMMT'25	-0.0122	-0.0065	-0.0019	-0.0016	-0.0002	-0.0003
BrUMO'25	-0.0033	-0.0042	-0.0063	-0.0006	0.0003	0.0005
All	-0.0049	-0.0040	-0.0041	-0.0015	0.0001	0.0007

The task-level deltas in Table 15 show the same small-sample trend with task-dependent strength. Full transfer lowers held-out MAE for every reported task at $n = 4$ and for all three utilities, with the largest discovery-side improvement on HMMT'25. By $n = 16$, most uniform and stability-skewed deltas are near zero or positive, indicating mild over-regularization after the top-p bank contains enough direct evidence.

Prior alignment. Greedy transfer improves held-out prediction when greedy correctness is aligned with the corresponding top-p row success probability. For each model–task pair, the alignment diagnostic compares the mean top-p success rate on questions solved by the greedy run with the mean top-p success rate on questions missed by the greedy run. It also records an extreme-conflict rate: greedy wrong with top-p row mean at least 0.75, or greedy correct with top-p row mean at most 0.25.

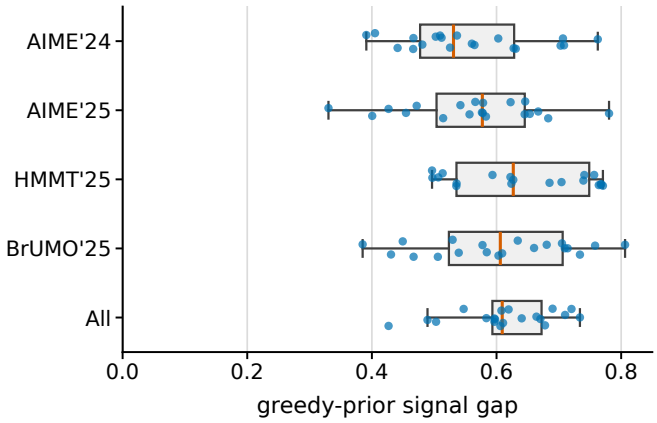

Figure 14: Distribution across models of the greedy-prior signal gap: mean top-p row success on greedy-correct questions minus mean top-p row success on greedy-wrong questions. Points are individual models; boxes show the interquartile range and median. Positive gaps indicate that the deterministic run is directionally informative about the stochastic bank. Nonzero conflict rates in Table 16 explain why large η can hurt some summaries.

The average signal gaps in Table 16 are positive on all reported tasks, so the greedy response contains information about which questions have high top-p success probability. The extreme-conflict rate ranges from 0.053 to 0.072 across tasks. These rows are where full transfer can pull posterior mass away from the top-p evidence, especially for higher-threshold utilities whose value is sensitive to reliable solvability under the stochastic decoding protocol.

Full-bank sensitivity. The optional prior has little effect once all 80 top-p trials are used as the observed bank. Table 17 reports the change from $\eta = 0$ to $\eta = 1$ for U_λ@16. Across the aggregate bank, the maximum absolute posterior-mean shift is 0.0027, and the average posterior standard-

Table 16: Greedy correctness as prior evidence for top-p success probabilities. The signal gap is the difference between the mean top-p row success rate on greedy-correct and greedy-wrong questions, averaged over models. The extreme-conflict rate is the fraction of questions where the greedy bit disagrees with a top-p row mean at threshold 0.75 or 0.25.

Task	Greedy acc.	Top-p mean	Signal gap	Extreme conflict
AIME'24	0.568	0.622	0.555	0.072
AIME'25	0.498	0.535	0.564	0.063
HMMT'25	0.312	0.305	0.645	0.053
BrUMO'25	0.555	0.597	0.604	0.060
All	0.483	0.515	0.615	0.062

deviation ratio is approximately 0.995. At the full bank size, the greedy prior has little influence on the final estimates of $U_\lambda(16)$.

Table 17: Sensitivity of the full 80-trial top-p posterior to full greedy-prior transfer ($\eta = 1$). Shifts are relative to the no-prior posterior for $U_\lambda@16$.

Task, λ	Mean $\lvert\Delta\mu\rvert$	Max $\lvert\Delta\mu\rvert$	Mean SD ratio
AIME'24, 0.5	0.0007	0.0017	0.995
AIME'24, 1	0.0009	0.0027	0.995
AIME'24, 2	0.0011	0.0042	0.995
AIME'25, 0.5	0.0008	0.0016	0.994
AIME'25, 1	0.0008	0.0015	0.995
AIME'25, 2	0.0009	0.0024	0.995
HMMT'25, 0.5	0.0009	0.0034	0.996
HMMT'25, 1	0.0006	0.0026	0.995
HMMT'25, 2	0.0006	0.0034	0.995
BrUMO'25, 0.5	0.0007	0.0018	0.994
BrUMO'25, 1	0.0009	0.0027	0.994
BrUMO'25, 2	0.0012	0.0037	0.995
All, 0.5	0.0006	0.0013	0.995
All, 1	0.0006	0.0018	0.995
All, 2	0.0007	0.0027	0.995

G.3.2 Budget Sensitivity of Greedy-prior Validation

The prior diagnostic in Section 4.3 fixes the prospective reporting budget at $k = 16$. The posterior update does not depend on that reporting budget; changing k changes the beta-binomial tail map in Section 3.1, the finite-bank hypergeometric reference in Equation (6), and the moment-family weights in Appendix C.8. The aggregate-bank validation is also evaluated at $k \in \{2, 4, 8\}$ using the paired protocol from Appendix J.3. The $k = 16$ rows in Table 18 reproduce the aggregate rows from Table 3 for comparison.

The smaller reporting budgets preserve the qualitative conclusion from the $k = 16$ diagnostic. With $n = 4$ observed top-p trials, full greedy transfer reduces held-out MAE for every tested k and λ (Table 18). For the discovery-skewed utility ($\lambda = .5$), ΔMAE ranges from -0.0039 at $k = 2$ to -0.0050 at $k = 8$, with -0.0049 at $k = 16$. For the uniform threshold average ($\lambda = 1$), the ΔMAE is -0.0040 at all four budgets up to rounding, consistent with the tail-sum identity relating the uniform profile average to single-sample accuracy in Section 3.2. For the repeatability-skewed utility ($\lambda = 2$), the small-sample improvement is similar across budgets, ranging from -0.0034 to -0.0041.

Once $n = 16$ stochastic trials are observed, the sign pattern matches the main diagnostic. Full greedy transfer still improves the discovery-skewed utility at every budget, but the gains are smaller, from -0.0004 at $k = 2$ to -0.0015 at $k = 16$. The uniform and repeatability-skewed utilities have positive ΔMAE at all four budgets. The prospective budget mainly changes effect size: the greedy prior is useful as low-sample evidence, especially for discovery-weighted summaries, and mild over-regularization remains visible for uniform and high-threshold summaries once the stochastic bank is already informative.

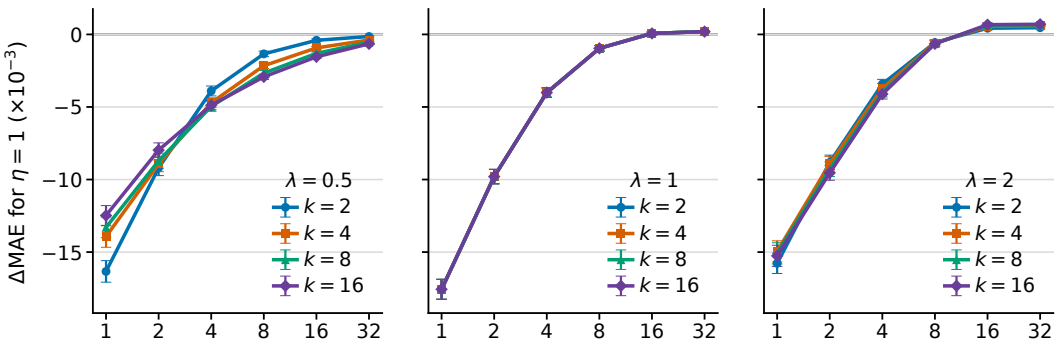

Figure 15: Budget sensitivity of full greedy-prior transfer on the aggregate response bank. Curves show paired changes in held-out MAE for $\eta = 1$ relative to the no-prior posterior across observed top-p sample sizes; negative values indicate improved prediction of the held-out finite-bank reference. Error bars are ± 1.96 standard errors over matched model–subsample units.

Table 18: Aggregate held-out validation across reporting budgets. Entries compare full greedy-prior transfer ($\eta = 1$) with the no-prior posterior ($\eta = 0$) for $U_\lambda @ k$ on the aggregate bank. Negative ΔMAE indicates lower error relative to the held-out top-p finite-bank reference.

		$n_{\text{obs}} = 4$			$n_{\text{obs}} = 16$		
k	λ	$\eta = 0$ MAE	$\eta = 1$ MAE	ΔMAE	$\eta = 0$ MAE	$\eta = 1$ MAE	ΔMAE
2	0.5	0.0398	0.0359	-0.0039	0.0149	0.0145	-0.0004
2	1	0.0429	0.0389	-0.0040	0.0152	0.0153	0.0001
2	2	0.0533	0.0499	-0.0034	0.0198	0.0203	0.0004
4	0.5	0.0467	0.0420	-0.0047	0.0194	0.0185	-0.0009
4	1	0.0429	0.0389	-0.0040	0.0152	0.0153	0.0001
4	2	0.0619	0.0582	-0.0037	0.0234	0.0240	0.0006
8	0.5	0.0544	0.0494	-0.0050	0.0250	0.0237	-0.0013
8	1	0.0429	0.0389	-0.0040	0.0152	0.0153	0.0001
8	2	0.0667	0.0627	-0.0040	0.0255	0.0261	0.0006
16	0.5	0.0590	0.0541	-0.0049	0.0299	0.0283	-0.0015
16	1	0.0429	0.0389	-0.0040	0.0152	0.0153	0.0001
16	2	0.0691	0.0650	-0.0041	0.0266	0.0273	0.0007

G.4 Categorical Evaluation

The categorical analysis applies the Dirichlet–categorical TailPass extension to a common set of 19 models across the four rubrics. The question set and sampling budget are fixed within these comparisons. The rubrics distinguish whether apparent success is concentrated in fully compliant, concise, or lower-perplexity correct attempts instead of only counting whether an exact answer appears. The categorical experiments in Section 4.4 use the protocol in Appendix J.6.

Table 19 gives the four finite schemata used in the experiments. The exact schema is the binary baseline. The format-aware schema separates boxed and unboxed outputs but assigns zero score to both wrong categories; its score vector discounts correct-but-unboxed attempts. The efficiency schema assigns partial credit to correct attempts according to completion length, and the confidence schema assigns partial credit to correct attempts according to completion perplexity. These two schemata use task-specific thresholds held fixed across models, trials, and questions within each task. Completion-length thresholds use correct attempts; perplexity thresholds use all attempts.

Table 20 reports the resulting task-level thresholds. Completion-length terciles are computed only among correct attempts, while wrong attempts receive zero credit regardless of length. Completion perplexity thresholds are computed over all attempts. The confidence-adjusted rubric assigns more credit to lower-perplexity correct completions. Interpreting these bands as correctness probabilities would require calibration against correctness labels. Pooled thresholds also make the rubric depend on the evaluated model cohort.

Table 19: Rubric-aware categorical schemata used in the $\mathrm{TailPass@}k$ experiments. Scores are normalized to $[0, 1]$ and are pre-specified before posterior aggregation.

Schema	Categories and scores	Interpretation
Exact	wrong=0.00; correct=1.00	Binary exact-answer correctness; the positive category is a correct final answer.
Format-aware	wrong/unboxed=0.00; wrong/boxed=0.00; correct/unboxed=0.70; correct/boxed=1.00	Correct boxed answers receive full credit; correct unboxed answers receive 0.70 because the final-answer protocol is not satisfied.
Efficiency-adjusted	wrong=0.00; correct/verbose=0.65; correct/moderate=0.85; correct/concise=1.00	Correct answers are discounted by completion length using task-level terciles among correct attempts.
Confidence-adjusted	wrong/high-conf=0.00; wrong/not-high-conf=0.00; correct/low-conf=0.70; correct/mid-conf=0.85; correct/high-conf=1.00	The confidence labels denote completion-perplexity bands: lower-perplexity correct answers receive more credit.

Table 20: Task-level thresholds used by the efficiency and confidence schemata. Token thresholds are completion-token terciles among correct attempts; confidence thresholds are completion-perplexity quantiles over all attempts. The confidence rubric assigns higher scores to lower-perplexity correct completions.

Task	Token $q_{.33}$	Token $q_{.66}$	PPL $q_{.40}$	PPL $q_{.60}$
AIME'24	5278	11020	1.156	1.217
AIME'25	4741	12303	1.161	1.224
BrUMO'25	4993	11058	1.153	1.209
HMMT'25	6621	13611	1.163	1.228

The Dirichlet$(1, \ldots, 1)$ prior assigns one count per category. Thus changing the category map changes both the scoring rule and the induced prior on correctness. These comparisons combine those two effects. The reported intervals condition on the estimated thresholds.

Table 21: Top models under aggregate rubric utility $U_{v,\lambda=2}(16)$, with v specified by the schema. Scores average posterior utility over the four tasks. Model short names are expanded in Table 35.

Schema	Rank	Model	$U_{v,\lambda=2}(16)$
Exact	1	Qwen3-Thinking	0.697 [0.692, 0.703]
Exact	2	Qwen3-4B	0.599 [0.592, 0.605]
Exact	3	Phi-4-plus	0.564 [0.557, 0.571]
Exact	4	gpt-oss-high	0.543 [0.535, 0.550]
Exact	5	AR-Nemotron	0.534 [0.527, 0.540]
Format-aware	1	Qwen3-Thinking	0.678 [0.671, 0.684]
Format-aware	2	Qwen3-4B	0.582 [0.575, 0.588]
Format-aware	3	Phi-4-plus	0.549 [0.542, 0.557]
Format-aware	4	gpt-oss-high	0.527 [0.519, 0.535]
Format-aware	5	AR-Nemotron	0.519 [0.511, 0.526]
Efficiency-adjusted	1	gpt-oss-medium	0.466 [0.459, 0.472]
Efficiency-adjusted	2	Phi-4	0.408 [0.402, 0.413]
Efficiency-adjusted	3	Phi-4-plus	0.403 [0.398, 0.408]
Efficiency-adjusted	4	gpt-oss-high	0.401 [0.396, 0.406]
Efficiency-adjusted	5	OpenThinker2	0.401 [0.396, 0.405]
Confidence-adjusted	1	Qwen3-Thinking	0.580 [0.574, 0.586]
Confidence-adjusted	2	Qwen3-4B	0.527 [0.520, 0.533]
Confidence-adjusted	3	AR-Nemotron	0.519 [0.512, 0.526]
Confidence-adjusted	4	OpenThinker2	0.470 [0.462, 0.477]
Confidence-adjusted	5	Light-R1-DS	0.464 [0.457, 0.471]

Table 21 reports the aggregate top-ranked models for the four categorical schemata at $k = 16$ and $\lambda = 2$. Exact and format-aware scoring produce the same top five, with lower displayed posterior means under the format-aware rubric. Efficiency-adjusted scoring changes the ordering: gpt-oss-medium and Phi-4 become the two leading models, while Qwen3-Thinking and Qwen3-4B move out of the top five. Confidence-adjusted scoring preserves Qwen3-Thinking and Qwen3-4B as the leading two models but promotes AR-Nemotron, indicating that exact-match performance and concentration in lower-perplexity correct categories need not rank systems identically.

Table 22: Largest rank changes induced by categorical rubrics relative to exact correctness, using aggregate rubric utility $U_{v,\lambda=2}(16)$. Positive Δ means that the rubric raises the model. Model short names are expanded in Table 35.

Model	Exact rank	Rubric rank	Δ
Format-aware			
AR-Nemotron	5	5	0
Bespoke	19	19	0
DS-R1-Qwen	18	18	0
EXAONE-4.0	15	15	0
FuseO1-DS-QwQ-SkyT1	8	8	0
LIMO-v2	11	11	0
Efficiency-adjusted			
Qwen3-4B	2	12	-10
Phi-4	10	2	8
Qwen3-Thinking	1	7	-6
gpt-oss-medium	6	1	5
AR-Nemotron	5	8	-3
FuseO1-DS-QwQ-SkyT1	8	6	2
Confidence-adjusted			
Phi-4-plus	3	9	-6
gpt-oss-medium	6	12	-6
gpt-oss-high	4	10	-6
OR1-Distill	13	8	5
LIMO-v2	11	6	5
EXAONE-4.0	15	11	4

Table 22 isolates the aggregate rank changes caused by moving from exact correctness to each categorical rubric at $k = 16$ and $\lambda = 2$. Format-aware scoring preserves the exact-match ranking in this comparison. Efficiency-adjusted scoring produces the largest shifts: Qwen3-4B moves down ten ranks and Qwen3-Thinking moves down six ranks, whereas Phi-4 and gpt-oss-medium move upward. Confidence-adjusted scoring gives a third view of the response bank; Phi-4-plus, gpt-oss-medium, and gpt-oss-high move down relative to exact correctness, while OR1-Distill, LIMO-v2, and EXAONE-4.0 move upward.

The task-level rankings in Table 23 show that the aggregate shifts appear across benchmarks. Qwen3-Thinking leads exact and format-aware utilities on all four tasks, but the efficiency rubric favors shorter correct completions on AIME'24, AIME'25, BrUMO'25, and HMMT'25. Under the confidence rubric, Qwen3-Thinking remains the leading aggregate model and AR-Nemotron appears in the top three on all four tasks, under a utility that weights correctness and completion perplexity jointly.

For one question, let $m_i = \sum_c \pi_{i,c} v_c$ and $s_i^2 = \sum_c \pi_{i,c} v_c^2 - m_i^2$. Conditional independence gives $\mathbb{E}[(A_{i,k}^v)^2 \mid \pi_i] = m_i^2 + s_i^2/k$, so the quadratic utility cannot increase with k when the posterior and rubric are fixed. Model rankings can nevertheless change because their m_i^2 and s_i^2 terms differ. In Figure 16, the relative ordering under the efficiency and confidence rubrics is stable over $k \geq 8$ for the leading models, supporting the primary-budget categorical comparison.

Figure 17 compares the same categorical rubrics over the moment-family spectrum. The discovery endpoint 0^+ rewards any positive rubric credit in the k-attempt bank. Increasing λ shifts mass toward high normalized bank scores and separates models by the concentration of high-value categories. The large efficiency and confidence shifts visible in the main text remain present for $\lambda = 1, 2$, and 4, indicating stability across moment parameters.

Table 23: Task-level top-three models under rubric utility $U_{v,\lambda=2}(16)$. Rows are grouped by benchmark and rank; entries give model short names followed by posterior mean utility. The columns specify rubric v: exact, format-aware, efficiency-adjusted, and confidence-adjusted scoring, respectively. Model short names are expanded in Table 35.

Task	Rank	Exact	Format	Efficiency	Confidence
AIME'24	1	Qwen3-Thinking (0.827)	Qwen3-Thinking (0.803)	gpt-oss-medium (0.580)	Qwen3-Thinking (0.676)
	2	Qwen3-4B (0.697)	Qwen3-4B (0.677)	gpt-oss-high (0.542)	AR-Nemotron (0.618)
	3	Phi-4-plus (0.680)	Phi-4-plus (0.661)	gpt-oss-low (0.520)	Qwen3-4B (0.601)
AIME'25	1	Qwen3-Thinking (0.741)	Qwen3-Thinking (0.720)	gpt-oss-medium (0.497)	Qwen3-Thinking (0.630)
	2	Qwen3-4B (0.663)	Qwen3-4B (0.643)	gpt-oss-high (0.470)	Qwen3-4B (0.600)
	3	gpt-oss-high (0.596)	gpt-oss-high (0.577)	gpt-oss-low (0.425)	AR-Nemotron (0.541)
BrUMO'25	1	Qwen3-Thinking (0.791)	Qwen3-Thinking (0.768)	OpenThinker2 (0.503)	Qwen3-Thinking (0.641)
	2	Qwen3-4B (0.679)	Qwen3-4B (0.660)	Phi-4 (0.498)	AR-Nemotron (0.624)
	3	OpenThinker2 (0.648)	OpenThinker2 (0.630)	gpt-oss-medium (0.492)	Qwen3-4B (0.582)
HMMT'25	1	Qwen3-Thinking (0.431)	Qwen3-Thinking (0.419)	gpt-oss-medium (0.294)	Qwen3-Thinking (0.374)
	2	Phi-4-plus (0.358)	Phi-4-plus (0.350)	Qwen3-Thinking (0.248)	Qwen3-4B (0.323)
	3	Qwen3-4B (0.356)	Qwen3-4B (0.347)	Phi-4-plus (0.241)	AR-Nemotron (0.293)

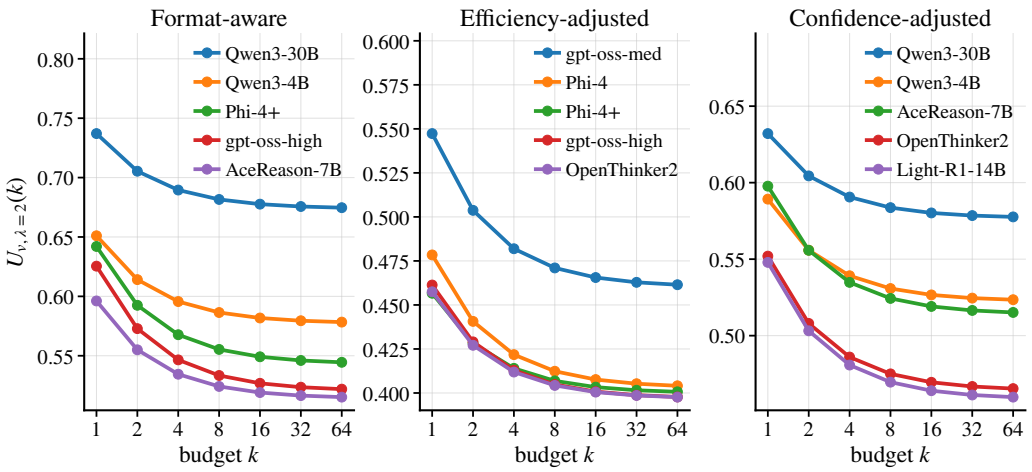

Figure 16: Budget dependence of aggregate categorical moment utilities $U_{v,\lambda=2}(k)$ for the indicated rubrics v. Each utility is nonincreasing in k under a fixed posterior and rubric, approaching the posterior mean of the latent squared per-attempt score.

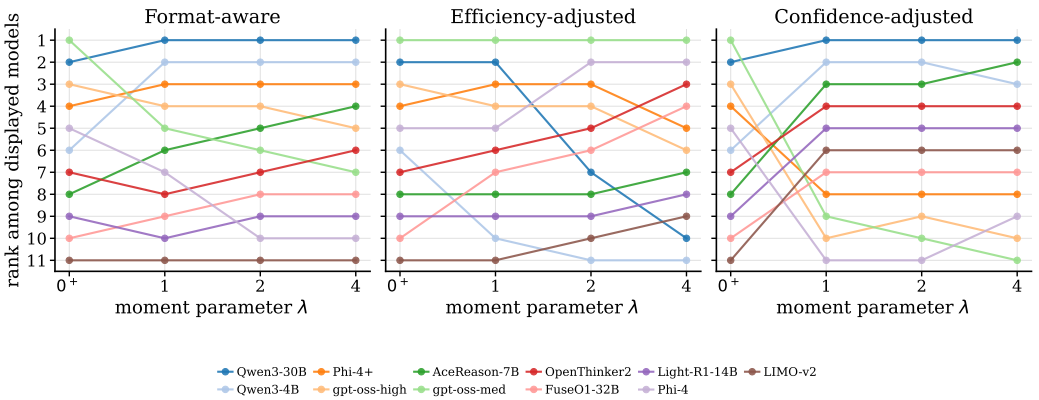

Figure 17: Rank sensitivity of categorical moment utilities $U_{v,\lambda}(16)$ over the moment exponent λ for the indicated rubrics v. The plotted models are the eleven highest-ranked models under exact correctness at $k = 16$ and $\lambda = 2$; ranks are recomputed within this displayed comparison set for each rubric and λ.

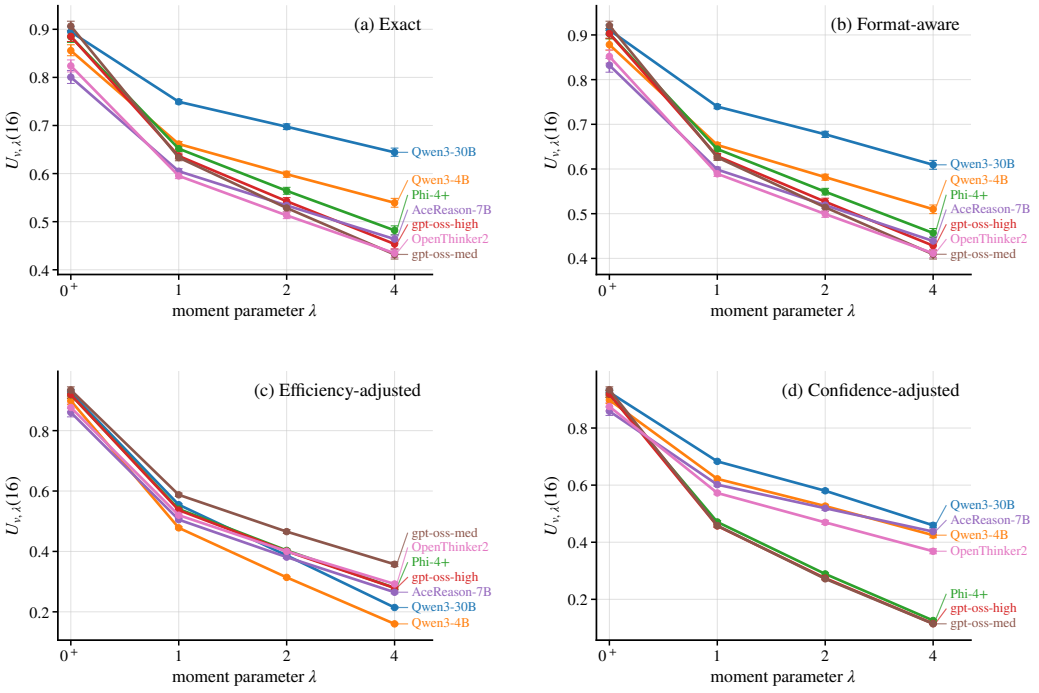

Figure 18: Posterior moment-utility spectra $U_{v,\lambda}(16)$ for the seven leading exact-match models, with rubric v specified by each panel. Points show posterior means and vertical bars show equal-tailed 95% credible intervals from shared Dirichlet posterior draws.

Figure 18 gives the posterior utility spectra for the leading exact-match models. Exact and format-aware spectra are nearly parallel, consistent with the small rank changes in Table 22. The efficiency spectra compress the high exact-match models and elevate systems with shorter correct completions. The confidence spectra separate models that have similar exact-match utility but different posterior mass on lower-perplexity correct attempts.

H Evaluation Across Mathematics and Science

TailPass separates discovery from repeatability through the required correct-response count. Moment utilities summarize the same fitted profile according to the value assigned to additional correct responses (Section 3). The Math and SuperGPQA banks exhibit both large repeatability gaps hidden by discovery scores and changes in the preferred configuration across these requirements.

H.1 Response banks and evaluation protocol

The Math and SuperGPQA response banks extend the evaluation to 186 competition-mathematics questions and 3,600 graduate-level questions (Hariri et al., 2026a; Hariri, 2026; M-A-P Team et al., 2025). Math comprises AIME 2026 (30 questions), CMIMC 2025 (40), HMMT February 2026 (33), HMMT November 2025 (30), and SMT 2025 (53). The SuperGPQA sample contains 50 questions in each of 72 fields across 13 disciplines. Both banks evaluate four configurations from two base models: Qwen3.6-35B-A3B and GPT-OSS-20B at high, medium, and low reasoning effort, with 80 recorded attempts per configuration and question. These give 59,520 Math attempts and 1,152,000 SuperGPQA attempts. The compact and detailed records for each bank represent the same attempts, which are counted once.

All configurations use temperature 1 and a maximum of 81,920 completion tokens. Qwen uses top-$p = 0.95$, top-$k = 20$, and presence penalty 1.5; GPT-OSS uses top-$p = 1$, no top-k truncation, and presence penalty zero. The configuration identity includes its reasoning effort and decoding settings. Every analysis retains the source correctness labels, including those for truncated or aborted attempts. Question identities and reference answers are aligned across configurations; the SuperGPQA joins use global question identifiers and UUIDs. Each bank contains exactly the seeds $0, \ldots, 79$, with no missing correctness labels or duplicated question–seed pairs.

Profiles use independent Beta$(1, 1)$ priors and exact posterior means from Equation (4). The primary reporting budget is $k = 16$, with descriptive profiles also evaluated at $k \in \{4, 8, 32\}$. All 80 attempts estimate the full-bank profiles. Equal-tailed 95% credible intervals use 20,000 joint posterior draws, shared across thresholds and moment utilities. These intervals condition on the sampled questions and the independent-attempt model. Question-level profiles are averaged within each benchmark or field; aggregate profiles weight questions equally. SuperGPQA held-out comparisons weight the 72 fields equally. Pointwise 95% performance-difference intervals use 4,000 paired bootstrap resamples of field means after averaging the reused partitions within each field (Efron, 1979). The field groups in Figure 19 summarize the observed Qwen bank; model-selection results use the individual fields and hold out attempts on the sampled questions. All results concern these recorded configurations and decoding settings.

H.2 Discovery, repeatability, and moment utilities

On AIME 2026, Qwen and GPT-OSS high have nearly identical discovery probabilities, $S_{16,1} = 0.9722$ and 0.9717, but their strict-repeatability probabilities are $S_{16,16} = 0.6766$ and 0.4793. The discovery difference is only 0.05 percentage points (pp), with a 95% posterior interval of $[-1.34, 1.42]$ pp. The strict-repeatability difference is 19.74 pp, with interval $[13.12, 26.28]$ pp. The intermediate thresholds distinguish the configurations well before the all-correct endpoint (Figure 20, top row). The repeatability-weighted utility $U_4(16)$ is 0.8282 for Qwen and 0.7303 for GPT-OSS high, a difference of 9.79 pp with interval $[6.97, 12.58]$ pp. Thus a near tie in discovery coexists with a well-separated repeatability utility.

SuperGPQA exhibits both hidden repeatability gaps and changes in the preferred configuration. Across its $72\binom{4}{2} = 432$ within-field pairs, 138 posterior mean profiles cross, 124 reverse between discovery and strict repeatability, and 44 reverse between $U_{.25}(16)$ and $U_4(16)$. Finite-bank pass@k rankings reverse with the budget $k = 1{:}80$ for 203 pairs. These are descriptive counts of strict point-estimate changes over all sampled fields.

Mechanics gives a pronounced discovery–repeatability reversal. GPT-OSS low has higher discovery than Qwen, 0.7948 versus 0.7029, while Qwen has higher strict repeatability, 0.3717 versus 0.0671. For Qwen minus GPT-OSS low, the respective difference intervals are $[-12.69, -5.55]$

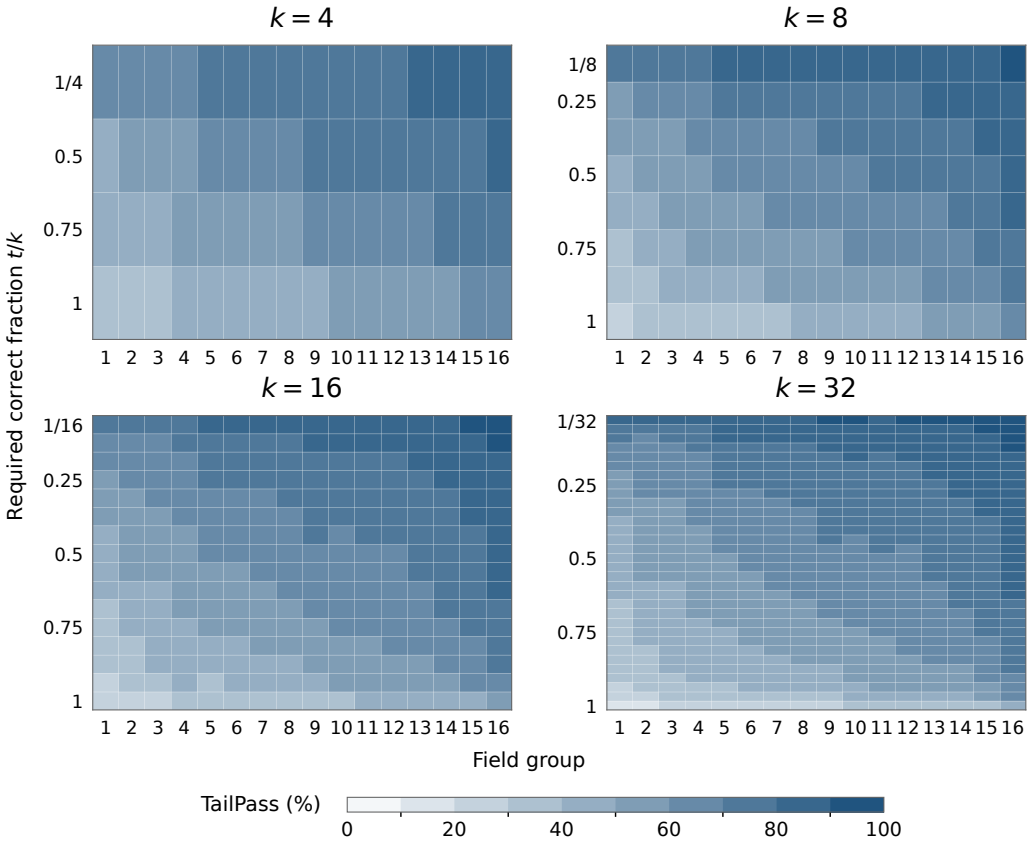

Figure 19: TailPass profiles for Qwen3.6-35B-A3B on the SuperGPQA response banks (M-A-P Team et al., 2025; Hariri, 2026). Panels show sampling budgets $k \in \{4, 8, 16, 32\}$. Columns are 16 groups (Table 24) formed by ordering the 72 sampled fields from lowest to highest observed accuracy and grouping consecutive fields in alternating sets of five and four; membership is fixed across budgets. Each field contains 50 questions with 80 recorded attempts per question. Rows give the required correct fraction t/k, increasing downward from $1/k$ to 1. Cell colors show the group-average posterior predictive probability of at least t correct responses in k fresh attempts, using independent question-level $\mathrm{Beta}(1, 1)$ priors. The first row measures discovery and the last requires every attempt to succeed.

and $[27.34, 33.33]$ pp. The full profile therefore separates broad occasional success from repeated correctness on the same questions.

Moment utilities make the preference change explicit. In Traditional Chinese Medicine, GPT-OSS low exceeds GPT-OSS high under the discovery-weighted $U_{.25}(16)$, 0.4995 versus 0.4644. The order reverses under $U_4(16)$, 0.0726 versus 0.1101. For high minus low, the difference intervals are $[-6.19, -0.89]$ pp at $\lambda = .25$ and $[2.57, 4.93]$ pp at $\lambda = 4$. Both utilities derive from the same fitted profile; only the declared marginal value of additional correct responses changes. The uniform member $U_1(16)$ equals posterior mean accuracy by Equation (10).

Joint posterior draws also distinguish a numerical ordering from a resolved comparison. A pair is unresolved when the equal-tailed 95% interval for its utility difference contains zero, following Appendix J.4. Table 25 shows how this assessment changes across the same profile. On Math, nine of the 30 discovery comparisons are unresolved, while all 30 are resolved for the six displayed moment utilities. On SuperGPQA, the unresolved count changes from 174 at discovery to 13 under $U_4(16)$. These counts apply a pointwise comparison rule; they do not assert simultaneous separation of all pairs. TailPass supplies both the utility-specific ordering and its posterior uncertainty. The number of resolved pairs depends on the utility; it is not a calibration test for the credible intervals.

Table 24: Field assignments for the 16 groups in Figure 19. Fields are ordered by Qwen3.6-35B-A3B's observed accuracy, with alphabetical tie-breaking. Each field contains 50 questions and 80 attempts per question. Accuracy is the observed correct fraction within each group.

Group	Assigned fields	Accuracy (%)
1	Journalism and Communication; Aquaculture; Art Studies; Atmospheric Science; Electrical Engineering	46.21
2	Animal Husbandry; Political Science; Veterinary Medicine; Naval Architecture and Ocean Engineering	51.13
3	Biology; Mechanics; Sociology; Library, Information and Archival Management; Musicology	53.12
4	Weapon Science and Technology; Psychology; Chemistry; Pharmacy	56.14
5	Stomatology; Petroleum and Natural Gas Engineering; Civil Engineering; Aeronautical and Astronautical Science and Technology; Forestry Engineering	59.35
6	Metallurgical Engineering; Geology; Public Administration; Agricultural Engineering	60.92
7	Public Health and Preventive Medicine; Business Administration; Military Science; Crop Science; Geography	62.11
8	History; Astronomy; Computer Science and Technology; Hydraulic Engineering	62.98
9	Forestry; Architecture; Language and Literature; Traditional Chinese Medicine; Physical Education	65.44
10	Theoretical Economics; Instrument Science and Technology; Materials Science and Engineering; Nuclear Science and Technology	66.99
11	Management Science and Engineering; Mining Engineering; Basic Medicine; Geological Resources and Geological Engineering; Physics	68.12
12	Transportation Engineering; Food Science and Engineering; Law; Environmental Science and Engineering	68.73
13	Optical Engineering; Geophysics; Philosophy; Textile Science and Engineering; Power Engineering and Engineering Thermophysics	70.97
14	Clinical Medicine; Electronic Science and Technology; Surveying and Mapping Science and Technology; Applied Economics	72.91
15	Systems Science; Mechanical Engineering; Physical Oceanography; Control Science and Engineering; Information and Communication Engineering	75.96
16	Education; Oceanography; Chemical Engineering and Technology; Mathematics	80.85

Table 25: Unresolved within-benchmark or within-field configuration pairs at $k = 16$, under the pairwise 95% posterior-interval rule. Every column uses the same response bank and shared posterior draws.

Dataset	Pairs	Discovery	$U_{.25}$	$U_{.5}$	U_1	U_2	U_4	U_8
Math	30	9	0	0	0	0	0	0
SuperGPQA	432	174	94	57	22	11	13	22

H.3 Threshold-specific model choice on SuperGPQA

Threshold-specific model choices are evaluated on disjoint SuperGPQA attempts, using the regret criterion from Appendix D.2. Each of 256 shared random permutations assigns the first $n \in \{16, 32, 48, 64\}$ attempts to fitting and the remaining $80 - n$ to evaluation, with nested fitting sets across n. All configurations, methods, thresholds, and utilities share these partitions. The reference $\widehat{S}^{\text{test}}_{m,16,t}$ is the exact hypergeometric probability of reaching threshold t in a 16-attempt subset of the evaluation bank (Equation (6)).

For a configuration pair $\{a, b\}$, let $\widehat{m}_{a,b,t}$ be the choice for threshold t. Regret within a field averages all six pairs and all 16 thresholds:

$$\mathcal{R} = \frac{1}{6 \cdot 16} \sum_{a<b} \sum_{t=1}^{16} \left[\max_{m \in \{a,b\}} \widehat{S}^{\text{test}}_{m,16,t} - \widehat{S}^{\text{test}}_{\widehat{m}_{a,b,t},16,t} \right]. \tag{42}$$

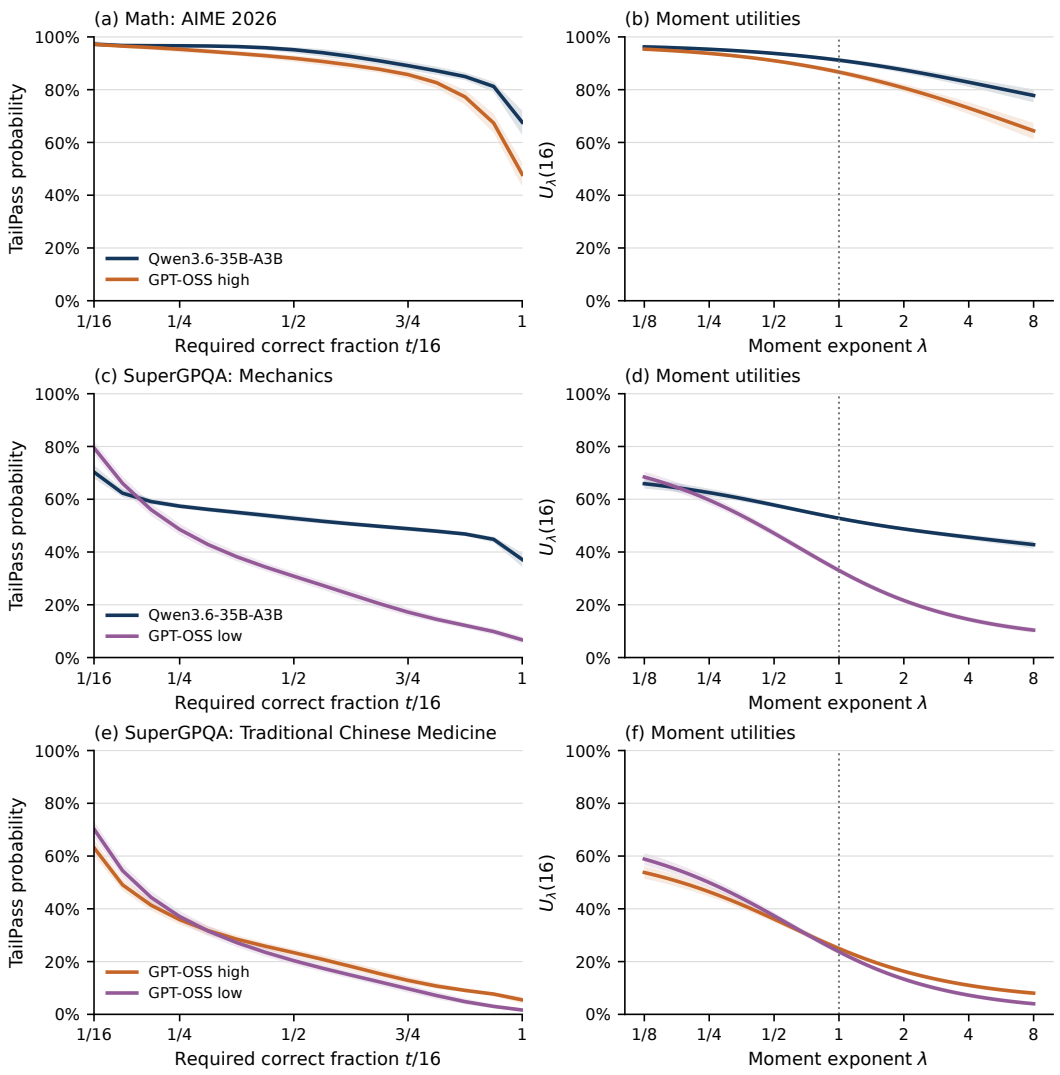

Figure 20: Profiles and moment utilities at $k = 16$ on the Math and SuperGPQA response banks. Left: posterior mean tail probabilities across correctness requirements. Right: the same profiles summarized by $U_\lambda(16)$; the dotted line marks uniform threshold weighting, $\lambda = 1$. Bands are pointwise 95% posterior intervals. AIME 2026 shows a discovery near tie with a large repeatability gap. Mechanics shows opposite endpoint preferences. Traditional Chinese Medicine shows a reversal within the moment family. These descriptive examples use fixed selection rules: the closest finite-bank discovery scores among Math pairs with a U_4 gap of at least five pp, the largest minimum opposing endpoint gap in SuperGPQA, and the largest minimum opposing $U_{.25}/U_4$ gap in SuperGPQA.

Tied fitting scores select uniformly among tied candidates. TailPass uses the posterior mean at the required threshold. The matched finite-bank profile uses the corresponding hypergeometric coordinate, equivalently G-Pass@16_τ at $\tau = t/16$ (Liu et al., 2025). The seven scalar baselines in Table 26 use one ordering across all thresholds, with definitions in Appendix M. Two binomial controls evaluate $q_{16,t}(\widehat{p}_i)$ at $\widehat{p}_i = c_i/n$ or $(c_i + 1)/(n + 2)$, where c_i is the correct fitting-response count for question i. The second shares TailPass's posterior mean success probability; integrating $q_{16,t}(p_i)$ over the posterior distinguishes TailPass from this plug-in estimate.

TailPass has the lowest mean threshold-selection regret among all 11 rules at every tested fitting size on SuperGPQA (Table 26). At $n = 16$, its regret is 0.112 pp, compared with 0.197 pp for the matched finite-bank profile, 0.118 pp for the binomial MLE, and 0.121 pp for the posterior-mean plug-in. The paired reductions are respectively 0.085 pp $[0.075, 0.095]$, 0.0056 pp $[0.0021, 0.0088]$, and 0.0084 pp $[0.0058, 0.0109]$. After averaging partitions within fields, TailPass improves on the MLE in 53 of

Table 26: SuperGPQA threshold-specific selection regret at $k = 16$ (pp; lower is better). Results average all 16 thresholds, six configuration pairs, and 256 partitions within each field, then weight the 72 fields equally. The final four rules condition on the required threshold; the others use a fixed scalar ordering. Fitting and evaluation use n and $80 - n$ disjoint attempts per question. Bold marks the lowest mean.

Selection rule	$n = 16$	$n = 32$	$n = 48$	$n = 64$
pass@16	2.633	2.468	2.472	2.542
pass^{16}	0.372	0.260	0.256	0.355
G-Pass@$16_{.5}$	0.346	0.261	0.265	0.359
Strict binary majority	0.320	0.244	0.249	0.350
mG-Pass@16	0.178	0.173	0.192	0.300
Discovery AUC	1.139	1.024	1.021	1.106
Accuracy	0.181	0.175	0.191	0.296
Binomial MLE	0.118	0.089	0.097	0.198
Binomial posterior mean	0.121	0.089	0.097	0.198
Matched finite-bank profile	0.197	0.100	0.102	0.202
TailPass profile	**0.112**	**0.086**	**0.095**	**0.197**

72 fields and on the posterior-mean plug-in in 58 fields. Relative to accuracy, the reduction is 0.069 pp $[0.040, 0.103]$.

H.4 Repeatability-weighted ranking on SuperGPQA

The repeatability-weighted utility $U_4(16)$ assigns increasing marginal value to additional correct responses. Its held-out target is

$$\widehat{U}_4^{\text{test}}(16) = \sum_{t=1}^{16} \omega_{t,16}^{(4)} \widehat{S}_{16,t}^{\text{test}}.$$

Kendall's τ_b measures agreement between fitting-score and held-out rankings, retaining ties (Kendall, 1945). The matched finite-bank moment applies the same weights to the fitting hypergeometric profile; TailPass-Moment applies them to the posterior profile. The binomial controls apply them to their plug-in profiles. Undefined correlations are omitted; at $n = 16$, these are 14 of 18,432 field–partition values for pass@16, with none for the other rules. Each fitting size uses its own $80 - n$-attempt evaluation bank.

Table 27: SuperGPQA ranking recovery for the repeatability-weighted utility $U_4(16)$ (mean held-out Kendall τ_b; higher is better). The final four rules estimate the same moment target. All methods share the fitting and held-out attempts; the 72 fields receive equal weight after averaging 256 partitions. Bold marks the highest mean.

Selection rule	$n = 16$	$n = 32$	$n = 48$	$n = 64$
pass@16	0.3961	0.3931	0.3942	0.3997
pass^{16}	0.8971	0.9327	0.9424	0.9423
G-Pass@$16_{.5}$	0.8728	0.9027	0.9097	0.9104
Strict binary majority	0.8889	0.9182	0.9251	0.9207
mG-Pass@16	0.9684	0.9791	0.9780	0.9676
Discovery AUC	0.6136	0.6320	0.6303	0.6269
Accuracy	0.9558	0.9712	0.9741	0.9651
Binomial MLE	0.9715	0.9811	0.9811	0.9717
Binomial posterior mean	0.9727	0.9816	0.9815	0.9720
Matched finite-bank moment	0.9702	0.9803	0.9805	0.9715
TailPass-Moment	**0.9735**	**0.9823**	**0.9819**	**0.9722**

TailPass-Moment has the highest mean ranking agreement for $U_4(16)$ at all four fitting sizes (Table 27). At $n = 16$, its $\tau_b = 0.9735$ exceeds 0.9702 for the matched finite-bank moment, 0.9715 for the binomial MLE, and 0.9727 for the posterior-mean plug-in. The paired gains over the binomial controls are 0.0020 $[0.0006, 0.0037]$ and 0.0008 $[0.00004, 0.00166]$. The common target and attempt partitions isolate ranking recovery for this repeatability utility.

I Empirical Response-bank Diagnostics for Scalar Repeated-Sampling Metrics

The empirical diagnostics are computed separately within each of the four available top-p response banks: AIME'24, AIME'25, HMMT'25, and BrUMO'25. Each bank has 20 models, $M = 30$ questions per model, and $N = 80$ sampled responses per question. Pairwise counts compare models within a fixed task, yielding $4\binom{20}{2} = 760$ task-specific model pairs before ties are removed. Finite-bank metrics use Equation (44). TailPass summaries use the posterior profile in Section 3; posterior intervals and pairwise probabilities are estimated from 5,000 joint posterior draws, with exact posterior means from Equation (4).

Table 28: Which failure modes are auditable from the binary response tensors. "Instantiated" means the failure can be computed directly from the tensors; "diagnosed" means the tensors show the scalar symptom but do not identify all latent causes.

Failure mode	Tensor status	Empirical symptom	TailPass consequence
Budget-induced $\mathrm{pass@}k$ rank instability	Instantiated	$\mathrm{pass@}k$ curves over $k = 1{:}80$ reverse for 138 of 760 task-specific model pairs.	Report the endpoint as discovery and compare it with the full profile or declared U_λ.
Large-budget target mismatch	Instantiated	$\mathrm{pass@}80$ is observed coverage; its strict ordering disagrees with average accuracy in 83 of 642 non-tied pairs.	The profile distinguishes the discovery and stability coordinates from its area.
Structural non-identifiability of a scalar	Diagnosed	Near-tied $\mathrm{pass@}16$ pairs can have visibly different high-threshold tails.	TailPass reports the tail curve and moment utilities rather than only one coordinate.
Threshold and area projections	Instantiated	G-Pass@$16_{0.5}$, majority, pass^{16}, mG-Pass@16, AUC, and U_λ induce different strict pairwise rankings.	Changing λ is a declared utility choice over the same estimated profile.
Finite-bank uncertainty	Instantiated	Several point-estimate winners lack 95% posterior support at $k = 16$.	Posterior intervals and pairwise probabilities accompany each profile coordinate and utility.
Modal-answer majority	Not identified	The tensors contain correctness bits but not answer clusters or tie-breaking information.	A modal-vote audit requires response strings or canonical answer classes.
Verifier and selector fragility	Not identified	The tensors do not separate true correctness from verifier acceptance or selected-candidate correctness.	The response record must include verifier decisions and selected outputs.
Harmful or costly attempts	Not identified	Incorrect responses are not categorized by harm, side effects, latency, or review cost.	A categorical outcome bank and explicit utility are required.
Attempt dependence and effective budget	Not identified	A single binary bank cannot distinguish sampling dependence from question difficulty heterogeneity.	Trace identifiers, duplicate diagnostics, or repeated independent batches are needed before correcting the budget axis.

Table 28 separates failures that the binary tensors can instantiate from failures that require richer response records. The instantiated cases are scalar-projection failures and finite-bank uncertainty. The non-identified cases are measurement failures: the correctness tensor does not contain answer clusters, verifier decisions, selected outputs, harm labels, latency or cost labels, trace identifiers, or repeated independent batches. A binary TailPass profile leaves missing labels unresolved and identifies which additional record is needed before the corresponding utility can be evaluated.

Table 29 reports the aggregate pairwise projection diagnostics. The largest disagreement with $\mathrm{pass@}16$ comes from the strict endpoint pass^{16}, which reverses 157 strict pairwise rankings. The stability-weighted moment utility $U_{\lambda=4}(16)$ reverses 137 rankings relative to $\mathrm{pass@}16$, while the discovery-weighted moment utility $U_{\lambda=.25}(16)$ reverses 42. The monotone increase in disagreement as λ moves toward high thresholds is consistent with the discovery–stability interpretation in Appendix C.8.

Table 29: Pairwise scalar-projection diagnostics across the four task-specific response banks. Counts are strict rank disagreements over task-specific model pairs; ties are omitted from the denominator.

Diagnostic	Count	Axis exposed by the disagreement
pass@k budget reversal over $k = 1{:}80$	138/760	budget axis
Accuracy versus pass@80	83/642	large-budget target
pass@16 versus G-Pass@$16_{0.5}$	98/760	threshold or utility projection
pass@16 versus Maj@16	107/760	threshold or utility projection
pass@16 versus pass^{16}	157/760	threshold or utility projection
pass@16 versus mG-Pass@16	129/760	threshold or utility projection
pass@16 versus AUC@16	30/760	threshold or utility projection
pass@16 versus $U_{\lambda=.25}(16)$	42/760	threshold or utility projection
pass@16 versus $U_{\lambda=1}(16)$	96/759	threshold or utility projection
pass@16 versus $U_{\lambda=4}(16)$	137/760	threshold or utility projection

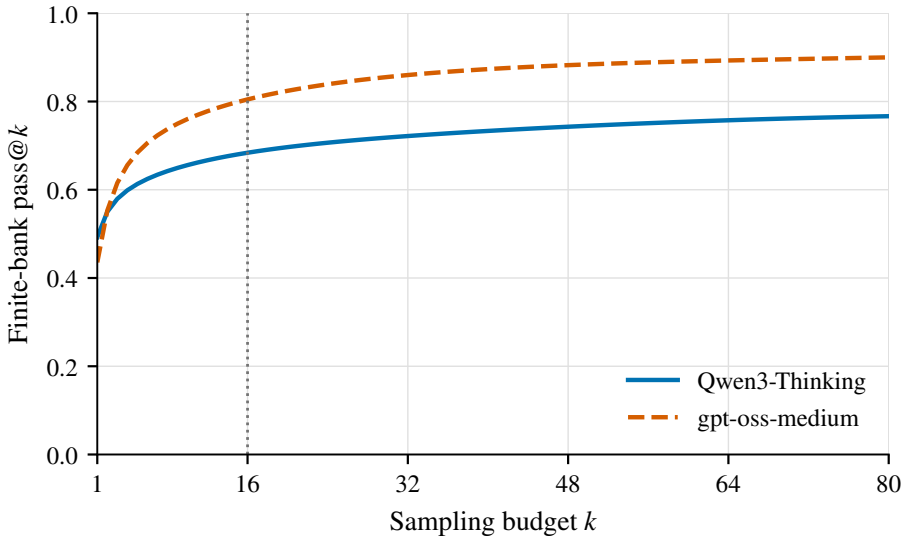

Figure 21: Budget-induced discovery reversal on HMMT'25. Qwen3-Thinking has the larger one-sample accuracy, but gpt-oss-medium has the larger finite-bank pass@k endpoint after the budget moves into the discovery regime. The dotted line marks the reporting budget $k = 16$.

Figure 21 shows the budget-axis component of the HMMT'25 pair highlighted in the main text. Qwen3-Thinking has higher single-sample accuracy, but gpt-oss-medium reaches more questions at least once as the discovery budget grows. This is the empirical analogue of the budget-induced rank instability in Appendix K.1.

Table 30: A near-alias of the pass@16 endpoint on HMMT'25. The two models have nearly identical discovery endpoints but different threshold profiles. Model short names are expanded in Table 35.

Score	OR-Nemotron	NVIDIA-Nemotron
pass@16	0.438	0.437
$U_{\lambda=.25}(16)$	0.348	0.390
$U_{\lambda=1}(16)$	0.236	0.311
$U_{\lambda=4}(16)$	0.149	0.213
Maj@16	0.239	0.322
pass^{16}	0.084	0.161

Table 30 and Figure 22 show a near tie of the pass@16 endpoint, 0.438 versus 0.437. Among HMMT'25 pairs with a finite-bank $U_{\lambda=4}(16)$ gap of at least 0.05, this pair has the smallest discovery-score gap. The rest of the threshold profile differs: NVIDIA-Nemotron has higher majority, higher strict stability, and higher moment utilities for all displayed λ values. Appendix K.3 gives the exact

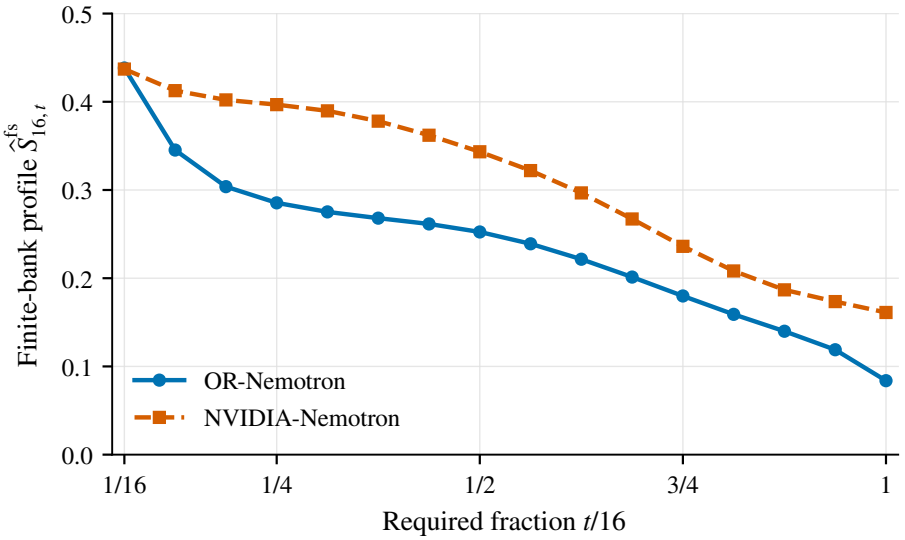

Figure 22: Near-tied pass@16 discovery endpoints with different tails on HMMT'25. Curves show the finite-bank profiles $\widehat{S}^{\mathrm{fs}}_{16,t}$ for OR-Nemotron and NVIDIA-Nemotron, with discovery at $t = 1$ and strict stability at $t = 16$. The discovery endpoints are nearly identical, but NVIDIA-Nemotron has a stronger intermediate and high-threshold tail.

structural non-identifiability result; this empirical pair shows an analogous scalar-aliasing symptom, where a single endpoint hides a large repeatability gap.

Table 31: Posterior ranking uncertainty for discovery $S_{16,1}$ and moment utilities $U_\lambda(16)$. Each entry counts task-specific model pairs whose posterior-mean winner has less than 95% posterior support under the stated score.

Task	$S_{16,1}$	$U_{\lambda=.25}(16)$	$U_{\lambda=1}(16)$	$U_{\lambda=4}(16)$
AIME'24	41	17	10	19
AIME'25	34	16	19	21
HMMT'25	41	30	9	17
BrUMO'25	35	24	13	12
Total	151	87	51	69

Table 31 reports posterior ranking uncertainty. The discovery endpoint has the largest number of comparisons with less than 95% support for the posterior-mean winner: 151 pairs, compared with 51 under uniform threshold utility. By the tail-sum identity, the uniform utility equals average single-sample accuracy under each posterior draw. The number of unresolved comparisons depends on both posterior uncertainty and the separation between model scores. Reporting the profile and its joint posterior makes these pairwise comparisons available across the full threshold spectrum.

I.1 Temperature Effects on the Discovery–Stability Spectrum

The experiments in Sections 4.1, 4.2, 4.3, and 4.4 use fixed decoding settings: top-p sampling for stochastic trials and greedy decoding for deterministic auxiliary runs. In the following experiment, we vary decoding temperature for Qwen3.5-4B on AIME'26 while holding the model and question set fixed, with 64 trials per question at each temperature. TailPass@k captures how the resulting response distributions differ across the discovery–stability spectrum.

Temperature sampling changes the response distribution at test time while the trained model remains fixed. Given next-token logits $z_v(h)$ at history h, temperature $T > 0$ rescales the next-token distribution (Holtzman et al., 2020):

$$q_T(v \mid h) = \frac{\exp(z_v(h)/T)}{\sum_u \exp(z_u(h)/T)}. \tag{43}$$

For this distribution, $T < 1$ sharpens the probabilities, $T > 1$ flattens them, and the limit $T \downarrow 0$ concentrates on the argmax when it is unique. This transformation compounds over an autoregressive trajectory, and question-level correctness can vary nonmonotonically with token entropy. In open-ended generation, beam search can produce repetitive or generic text. Top-k sampling restricts each draw to the k most likely tokens (Fan et al., 2018), while nucleus sampling retains a variable number of tokens whose cumulative probability reaches a chosen threshold (Holtzman et al., 2020). In reasoning and code settings, the same sampled response distribution becomes an axis of test-time scaling: $\mathrm{pass@}k$ measures whether any of k attempts solves a task (Chen et al., 2021); self-consistency samples multiple reasoning paths before aggregating answers (Wang et al., 2023); repeated sampling studies report coverage gains as the number of attempts increases (Brown et al., 2024); and compute-optimal test-time scaling studies compare best-of-N, verifier-guided search, and adaptive allocation of inference compute across problem instances (Snell et al., 2025).

Temperature changes the task-level response distribution that induces the latent correctness probabilities in the TailPass profile. The estimand and posterior computation follow the binary $\mathrm{TailPass@}k$ formulation in Section 3, using the Beta–Bernoulli posterior from Section 3.1. Scalar summaries use the moment-family utilities in Appendix C.8 and the question-first $L_2@k$ spectrum score in Appendix C.8. The sweep protocol, posterior draw count, and exclusion rule for all-zero high-temperature response banks are specified in Appendix J.2.

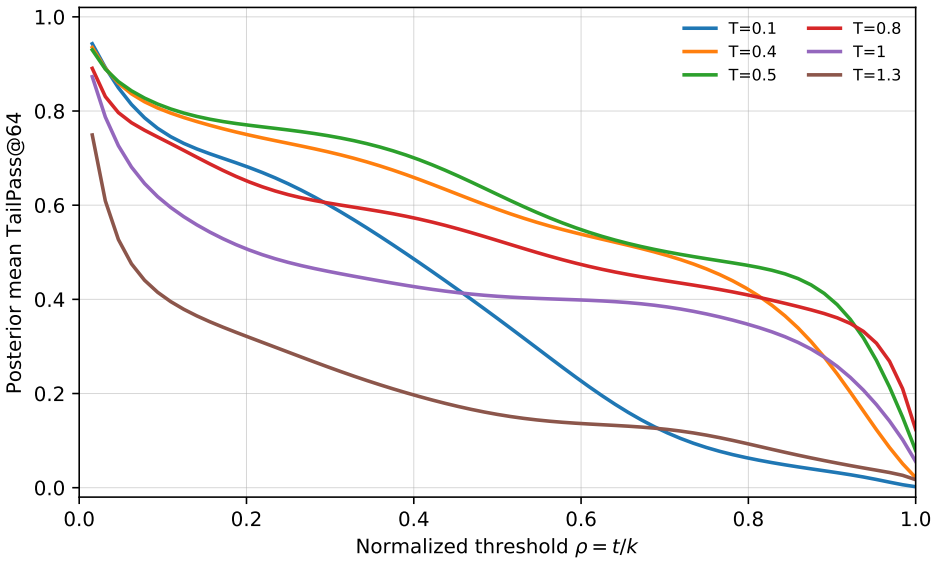

Figure 23: Posterior mean $\mathrm{TailPass@64}$ profiles for selected temperatures with at least one scored correct response. The x-axis is the normalized threshold $\rho = t/64$. Low-threshold coordinates measure discovery, high-threshold coordinates measure stability, and the interior shows how quickly correctness frequency decays as the required number of correct attempts increases.

The observed trial-level accuracy peaks at $T = 0.5$ in the nonzero range, and its posterior uniform-threshold counterpart $U_{\lambda=1}$ peaks at the same temperature. The $\mathrm{TailPass@64}$ curves reveal substantial threshold variation. At $T = 0.1$, discovery at $k = 64$ is high, 0.943 with 95% interval $[0.906, 0.976]$, but the curve collapses rapidly: the midpoint coordinate is 0.359 and the all-correct stability endpoint is 0.002. Temperatures 0.4–0.7 keep high discovery while raising much of the interior profile, and $T = 0.5$ gives the largest $L_2@64$ and the largest moment-utility values in Table 33. The stability endpoint alone is largest near $T = 0.8$–0.9, but those temperatures have lower midpoint, uniform-area, and L_2 summaries than $T = 0.5$. A stability-only endpoint can therefore favor a regime that is weaker over most of the threshold spectrum.

Table 34 gives the extrema among the retained temperatures. Discovery has its largest posterior mean at $T = 0.2$, $S_{64,1} = 0.946$ with interval $[0.910, 0.978]$, but all-correct stability is nearly zero, $S_{64,64} = 0.001$ with interval $[0.000, 0.006]$. All-correct stability is largest at $T = 0.8$, $S_{64,64} = 0.124$ with interval $[0.075, 0.178]$, although the same run has lower midpoint (0.525), uniform-area

Table 32: Profile coordinates $S_{64,t}$ and spectrum score L_2@64 for selected temperatures. Entries are posterior means with 95% equal-tailed intervals in brackets. "Scored acc." is the empirical trial-level accuracy; the remaining columns are posterior summaries of the TailPass@64 profile from the binary correctness matrices.

T	Scored acc.	Discovery $S_{64,1}$	Midpoint $S_{64,32}$	Stability $S_{64,64}$	L_2@64
0.1	0.368	0.943 [0.906, 0.976]	0.359 [0.308, 0.408]	0.002 [0.000, 0.011]	0.525 [0.507, 0.543]
0.3	0.525	0.926 [0.887, 0.967]	0.580 [0.549, 0.614]	0.001 [0.000, 0.007]	0.637 [0.622, 0.652]
0.5	0.607	0.930 [0.893, 0.968]	0.623 [0.582, 0.665]	0.079 [0.039, 0.121]	0.690 [0.676, 0.704]
0.7	0.558	0.920 [0.878, 0.960]	0.575 [0.541, 0.606]	0.112 [0.064, 0.163]	0.647 [0.632, 0.661]
0.9	0.476	0.888 [0.839, 0.937]	0.440 [0.413, 0.465]	0.122 [0.072, 0.174]	0.567 [0.552, 0.583]
1.1	0.354	0.802 [0.740, 0.865]	0.383 [0.359, 0.399]	0.038 [0.012, 0.072]	0.447 [0.430, 0.464]
1.3	0.195	0.750 [0.677, 0.822]	0.156 [0.136, 0.183]	0.017 [0.001, 0.033]	0.311 [0.292, 0.329]

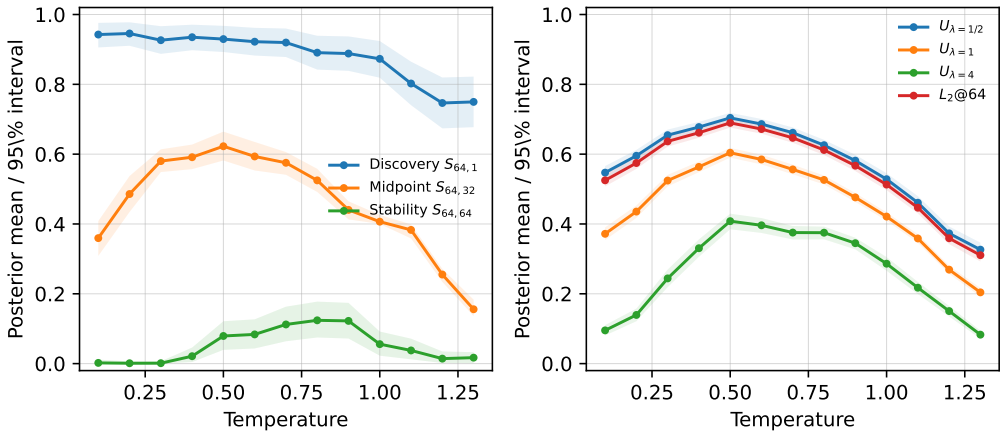

Figure 24: Temperature dependence of TailPass@64 endpoints and spectrum summaries for $T = 0.1$–1.3, the range with at least one scored correct response. Shaded bands are pointwise 95% equal-tailed posterior intervals. The right panel uses the moment-family weights in Equation (29); $\lambda = 1/2$ is more discovery-weighted, $\lambda = 1$ is the uniform threshold area, and $\lambda = 4$ is more stability-weighted.

Table 33: Moment-family expected-utility summaries at $k = 64$ for selected temperatures. Entries are posterior means with 95% equal-tailed intervals in brackets. The same TailPass@64 profile can support different rankings when the operational viewpoint shifts from discovery-heavy to stability-heavy thresholds.

T	$U_{\lambda=1/2}$	$U_{\lambda=1}$	$U_{\lambda=4}$
0.1	0.547 [0.529, 0.566]	0.372 [0.354, 0.390]	0.095 [0.081, 0.110]
0.3	0.654 [0.639, 0.670]	0.524 [0.508, 0.541]	0.244 [0.219, 0.269]
0.5	0.704 [0.690, 0.718]	0.604 [0.590, 0.618]	0.408 [0.385, 0.430]
0.7	0.661 [0.647, 0.677]	0.556 [0.542, 0.570]	0.375 [0.356, 0.394]
0.9	0.581 [0.565, 0.598]	0.476 [0.463, 0.489]	0.345 [0.327, 0.361]
1.1	0.461 [0.443, 0.480]	0.359 [0.346, 0.371]	0.217 [0.199, 0.234]
1.3	0.326 [0.307, 0.346]	0.204 [0.192, 0.217]	0.083 [0.071, 0.094]

Table 34: Posterior-mean extrema over the retained temperature range $T = 0.1, \ldots, 1.3$ at $k = 64$. Intervals are 95% equal-tailed posterior intervals, except for the scored trial accuracy, which is empirical. The extrema use analytic posterior means for linear summaries and Monte Carlo means for L_2@64.

Quantity	Posterior-mean maximizer	Value at maximizer
Scored trial accuracy	$T = 0.5$	0.607
Discovery endpoint $S_{64,1}$	$T = 0.2$	0.946 $[0.910, 0.978]$
Midpoint coordinate $S_{64,32}$	$T = 0.5$	0.623 $[0.582, 0.665]$
All-correct stability endpoint $S_{64,64}$	$T = 0.8$	0.124 $[0.075, 0.178]$
Discovery-weighted utility $U_{\lambda=1/2}$	$T = 0.5$	0.704 $[0.690, 0.718]$
Uniform threshold utility $U_{\lambda=1}$	$T = 0.5$	0.604 $[0.590, 0.618]$
Stability-weighted utility $U_{\lambda=4}$	$T = 0.5$	0.408 $[0.385, 0.430]$
Quadratic RMS score L_2@64	$T = 0.5$	0.690 $[0.676, 0.704]$

(0.526), and L_2@64 (0.612) summaries than $T = 0.5$. The largest posterior means for the reported moment utilities and L_2@64 occur at $T = 0.5$: scored trial accuracy is 0.607, $S_{64,32} = 0.623$ with interval $[0.582, 0.665]$, $U_{\lambda=1} = 0.604$ with interval $[0.590, 0.618]$, $U_{\lambda=1/2} = 0.704$ with interval $[0.690, 0.718]$, $U_{\lambda=4} = 0.408$ with interval $[0.385, 0.430]$, and L_2@64 $= 0.690$ with interval $[0.676, 0.704]$. These extrema compare posterior means; uncertainty in pairwise orderings requires the posterior distribution of the corresponding differences.

By retaining all thresholds, TailPass@k separates these operating points in one profile. Within $T = 0.1, \ldots, 1.3$, discovery has its largest posterior mean at $T = 0.2$, all-correct stability at $T = 0.8$, and the reported moment utilities at $T = 0.5$. pass@64 reports the discovery endpoint alone (Chen et al., 2021). The full profile makes the different temperature preferences across discovery and repeatability visible. The excluded temperatures $T = 1.4, \ldots, 2.2$ have all-zero correctness labels under the answer extractor (Appendix J.2).

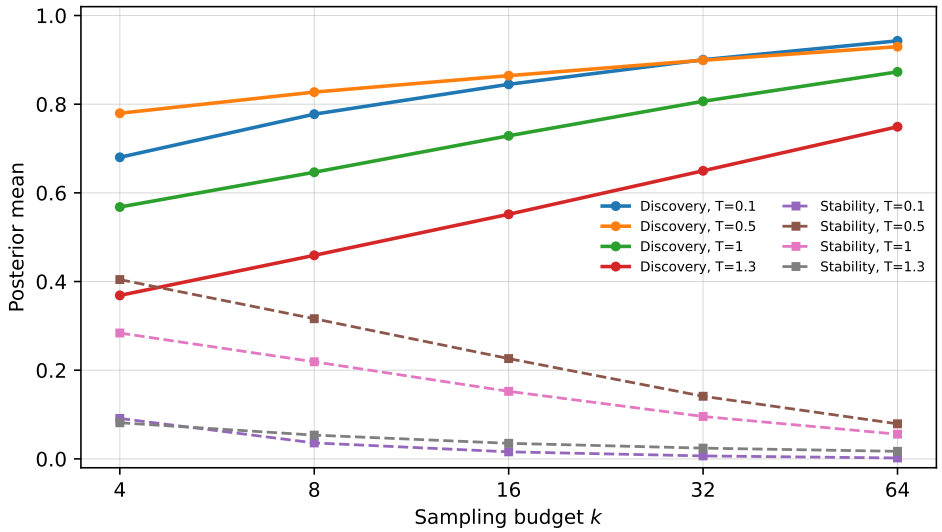

Figure 25: Endpoint behavior as the reporting budget k changes. For a fixed per-question success probability, discovery is nondecreasing in k while the all-correct stability endpoint is nonincreasing in k. The empirical posterior summaries show this definitional tradeoff and its interaction with temperature.

The budget sweep in Figure 25 gives the same conclusion from the test-time scaling viewpoint. Increasing k expands discovery because more attempts give more chances to see at least one correct answer, matching the coverage-oriented interpretation of repeated sampling (Brown et al., 2024). The same increase in k makes the all-correct stability event more stringent. At $T = 0.5$, posterior mean

discovery rises from 0.780 at $k = 4$ to 0.930 at $k = 64$, while stability falls from 0.404 to 0.079. At $T = 0.1$, discovery rises from 0.680 to 0.943, but stability falls from 0.091 to 0.002. $\mathrm{pass@}k$ records only the first of these movements. $\mathrm{TailPass@}k$ records the whole threshold profile and makes the sampling-regime tradeoff visible.

J Experimental Setup & Reproducibility

J.1 Models & Datasets

Datasets. The 20-configuration TTS response banks use four Olympiad-style mathematics benchmarks: AIME'24 (MathArena, 2025a;c), AIME'25 (MathArena, 2025b), BrUMO'25 (Brown University Math Olympiad Organizers, 2025), and HMMT'25 (Harvard–MIT Mathematics Tournament, 2025). For AIME'24 and AIME'25, AIME I and AIME II are combined within each year, giving 30 integer-answer problems per benchmark. The Qwen3.5-4B temperature-sweep experiment uses AIME'26 (MathArena, 2026), again combining AIME I and AIME II into $M = 30$ integer-answer problems. The HMMT'25 benchmark uses the official February 2025 contest set, covering algebra, geometry, number theory, and combinatorics. The BrUMO'25 benchmark uses the published 2025 problem sets from the tournament archive.

Models. The base cohort contains 11 evaluated configurations: eight non-`gpt-oss` reasoning checkpoints and the low, medium, and high reasoning-effort settings of `gpt-oss-20b` (OpenAI, 2025). The eight non-`gpt-oss` checkpoints are Sky-T1-32B-Flash (Griggs et al., 2025), Qwen3-30B-A3B-Thinking-2507 (Qwen, 2025a), DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025), LIMO-v2 (Ye et al., 2025), EXAONE-4.0-1.2B (Bae et al., 2025), OpenReasoning-Nemotron-1.5B (NVIDIA, 2025b), OpenThinker2-32B (Guha et al., 2025), and OpenThinker3-1.5B (OpenThoughts, 2025). A larger cohort adds nine reasoning-capable models for sensitivity to the set of evaluated systems: Phi-4-reasoning and Phi-4-reasoning-plus (Abdin et al., 2025), OpenR1-Distill-7B (Hugging Face, 2025), FuseO1-DeepSeekR1-QwQ-SkyT1-Flash-32B-Preview (FuseAI, 2025), Light-R1-14B-DS (Wen et al., 2025a), AceReason-Nemotron-1.1-7B (Liu et al., 2026), NVIDIA-Nemotron-Nano-9B-v2 (NVIDIA, 2025a), Qwen3-4B-Thinking-2507 (Qwen, 2025b), and Bespoke-Stratos-7B (Bespoke Labs, 2025). Table 35 gives the model identifiers used in figures and result tables.

Table 35: Model identifiers used in figures and result tables; the short-name column gives the notation used elsewhere.

ID	Model	Short name
1	DeepSeek-R1-Distill-Qwen-1.5B	DS-R1-Qwen
2	LIMO-v2	LIMO-v2
3	OpenThinker2-32B	OpenThinker2
4	OpenThinker3-1.5B	OpenThinker3
5	Qwen3-30B-A3B-Thinking-2507	Qwen3-Thinking
6	Sky-T1-32B-Flash	Sky-T1-Flash
7	gpt-oss-20b_high	gpt-oss-high
8	gpt-oss-20b_low	gpt-oss-low
9	gpt-oss-20b_medium	gpt-oss-medium
10	EXAONE-4.0-1.2B	EXAONE-4.0
11	OpenReasoning-Nemotron-1.5B	OR-Nemotron
12	Phi-4-reasoning	Phi-4
13	Phi-4-reasoning-plus	Phi-4-plus
14	OpenR1-Distill-7B	OR1-Distill
15	FuseO1-DeepSeekR1-QwQ-SkyT1-Flash-32B-Preview	FuseO1-DS-QwQ-SkyT1
16	Light-R1-14B-DS	Light-R1-DS
17	AceReason-Nemotron-1.1-7B	AR-Nemotron
18	NVIDIA-Nemotron-Nano-9B-v2	NVIDIA-Nemotron
19	Qwen3-4B-Thinking-2507	Qwen3-4B
20	Bespoke-Stratos-7B	Bespoke

Prompting. Prompt templates follow the model documentation when a model-specific template is available; otherwise, the evaluation uses the instruction, "Please reason step by step, and put your final answer within `\boxed{}`." For `gpt-oss-20b`, the Harmony prompt template specifies three discrete reasoning-effort levels. For `OpenReasoning-Nemotron-1.5B`, the task instruction is "Solve the following math problem. Make sure to put the answer (and only the answer) inside `\boxed{}`."

J.2 Sampling Protocols

The 20-configuration TTS response banks use top-p sampling with temperature 0.6, $p = 0.95$, batch size 1, and random seeds 1234 through 1313, yielding $N = 80$ trials for each dataset–model pair. For these banks, the task set is AIME'24, AIME'25, HMMT'25, and BrUMO'25. Each task tensor has shape $20 \times 30 \times 80$: 20 model configurations, $M = 30$ questions, and $N = 80$ stochastic trials per question.

The Qwen3.5-4B temperature sweep on AIME'26 covers temperatures $0.1, 0.2, \ldots, 2.2$. For each temperature, the response bank contains $M = 30$ questions and $N = 64$ scored trials per question. The analysis treats the trials within a temperature as exchangeable attempts from that temperature's response distribution, uses the Beta$(1, 1)$ no-prior baseline from Section 3.1, and reports posterior predictive $\mathrm{TailPass@64}$ summaries unless otherwise stated. Credible intervals in the tables and bands in the figures are equal-tailed posterior intervals estimated from $H = 5000$ joint posterior draws.

The response matrices for $T = 1.4, 1.5, \ldots, 2.2$ each contain zero scored correct answers out of 1920 trials under the answer extractor used for this response bank. These nine banks induce identical posterior profile distributions under the common prior. They are not included in the $\mathrm{TailPass@64}$ plots or scalar tables reported for the range with observed successes. With the Beta$(1, 1)$ prior, 64 failures give a Beta$(1, 65)$ posterior and predictive discovery probability $64/129$ at $k = 64$. This positive value reflects posterior uncertainty after finite evidence. Distinguishing incorrect answers from extraction errors requires review of the generated responses. All plotted and tabulated $\mathrm{TailPass@64}$ results are restricted to $T = 0.1, \ldots, 1.3$, where at least one scored correct answer is present.

J.3 Prior-Evidence Validation Protocol

The prior-evidence experiment uses the top-p response tensor from Appendix J.2 as the observed stochastic bank R and the single greedy decode as the auxiliary bank R^0 with $D = 1$. The aggregate averages the same 120 questions equally for all 20 model configurations. The no-prior baseline sets $D = 0$, equivalently $\eta = 0$. Transfer weights $\eta \in \{0, .25, .5, 1\}$ range from no transfer to a full one-trial greedy prior. Greedy outcomes enter only through the Beta prior parameters in Equation (3). The scalar summaries are the linear moment-family utilities $U_\lambda@k$ from Appendix C.8, with $\lambda = .5$ weighting low discovery thresholds, $\lambda = 1$ giving the uniform threshold average, and $\lambda = 2$ weighting higher repeatability thresholds. The main validation fixes $k = 16$; Appendix G.3.2 reports $k \in \{2, 4, 8, 16\}$.

For each model and benchmark, $n \in \{1, 2, 4, 8, 16, 32\}$ top-p trials per question are sampled without replacement and treated as observed evidence. The remaining $80 - n$ top-p trials form a held-out finite-bank reference, computed by applying the same moment-family utility to the hypergeometric finite-sample profile in Equation (6). Each condition uses 200 repetitions. For a fixed n and repetition, the same top-p trial indices are used for all models, moment-family utilities, and η values, so ΔMAE is a paired comparison with shared subsampling noise. The aggregate bank is the main validation bank; task-level analyses repeat the protocol on AIME'24, AIME'25, HMMT'25, and BrUMO'25.

Posterior means and exact posterior variances of the scalar utilities are computed from the linear-utility covariance formula in Equation (27). Reported 95% intervals are the clipped normal moment intervals from Appendix C.4. The held-out finite-bank value is a noisy reference for the latent utility. The reported coverages are containment rates for this realized reference. Predictive intervals for a newly realized benchmark would also include the finite-sampling layer from Appendix C.1.

J.4 TailPass@8 Uncertainty Protocol

The TailPass@8 uncertainty experiments use the top-p stochastic response bank from Appendix J.2. Each model–task slice is treated as a binary matrix with questions on the first axis and 80 sampled trials on the second axis; task-specific matrices have $M = 30$ questions, and the aggregate bank has $M = 120$ questions. For question i, the observed count γ_i updates the no-prior baseline from Section 3.1,

$$p_i \mid R_i \sim \mathrm{Beta}(1 + \gamma_i, 1 + 80 - \gamma_i),$$

equivalently $D = 0$ with an independent uniform $\mathrm{Beta}(1, 1)$ prior. All reported intervals are equal-tailed 95% posterior credible intervals estimated from $H = 20{,}000$ joint posterior draws.

For each model, a joint draw samples all question-level probabilities and then evaluates every TailPass@8 profile coordinate on that same draw. Scalar moment-family utilities $U_\lambda(8)$ are applied to the resulting profile draws, so profile coordinates and scalar summaries share the same posterior simulation. This is the dependence-preserving propagation described in Appendices C.3 and C.8.

The primary ranking view uses $U_{\lambda=2}(8)$. For each task, the strict reference ranking orders models by posterior mean under this same scalar view. For a model pair (a, b), posterior draws of $U^a_\lambda(8) - U^b_\lambda(8)$ define the pairwise comparison. The pair is unresolved when the equal-tailed 95% credible interval for the difference contains zero. Models are sorted by posterior mean, adjacent unresolved comparisons share a dense rank, and larger displayed groups are the transitive closure of adjacent unresolved comparisons. The all-pairs unresolved count is computed before this adjacent grouping step. A group can contain pairs whose difference intervals exclude zero. The 95% level applies to individual pairwise intervals; grouping follows the transitive adjacency rule. Because the uncertainty-aware rankings contain ties, agreement with the strict reference ranking is reported using Kendall's τ_b (Kendall, 1945). The scalar-view sensitivity experiment repeats the same procedure on the aggregate bank for $\lambda \in \{0.25, 0.5, 1, 2, 4, 8\}$ while holding the response bank, posterior, budget, and posterior draws fixed.

J.5 Response-bank Diagnostic Protocol

The empirical scalar-metric diagnostics use the four task-specific top-p response banks: AIME'24, AIME'25, HMMT'25, and BrUMO'25. Each task contains 20 models, 30 questions per model, and 80 sampled responses per model–question pair, yielding $4\binom{20}{2} = 760$ task-specific model pairs for pairwise comparisons before ties are removed. Finite-bank baselines are computed from the hypergeometric profile in Equation (44). The scalar baselines are $\mathrm{pass@}k$, $\text{G-Pass@}k_{0.5}$, $\mathrm{Maj@}k$, pass^k, $\text{mG-Pass@}k$, and $\mathrm{AUC@}K$, matching the projections in Appendix M. In these diagnostics, $\mathrm{AUC@}K$ assigns weight $1/K$ to each discovery budget $1, \ldots, K$. $\mathrm{TailPass@}k$ uses the posterior profile from Section 3 with the uniform $\mathrm{Beta}(1, 1)$ prior and the moment-family utilities from Appendix C.8. Unless otherwise stated, the reporting budget for these diagnostics is $k = 16$.

J.6 Categorical Evaluation Protocol

The categorical experiments use the top-p responses, retaining 19 models, 80 trials, and 30 questions per task. The preprocessing excludes NVIDIA-Nemotron because of nonfinite prompt log-probability fields, and the same 19-model cohort is used for all four rubrics. Binary evaluation retains this model. The completion-perplexity field used by the confidence rubric is finite for all 20 models in the release. A model's aggregate categorical score averages over all task-question pairs, so each of the 120 question-level latent rubric distributions receives equal weight.

Four pre-specified schemata map attempts to categorical labels: exact correctness, format-aware correctness, efficiency-adjusted correctness, and confidence-adjusted correctness. Format-aware scoring discounts correct answers without a boxed final answer, efficiency-adjusted scoring discounts correct but verbose completions using task-level completion-token terciles, and confidence-adjusted scoring discounts correct answers with higher completion perplexity. The posterior uses a uniform Dirichlet prior, no transfer prior, and equal-tailed 95% credible intervals from 1000 shared posterior draws. The primary categorical reporting budget is $k = 16$; budget sweeps use $k \in \{1, 2, 4, 8, 16, 32, 64\}$. Scalar rankings use the moment-family utility $U_{v,\lambda}(k)$ from Section 3.3 and Appendix C.7, with $\lambda = 2$ as the primary stability-sensitive setting and 0^+, 1, and 4 as sensitivity parameters.

J.7 Reproducibility

For the 20-configuration TTS banks, models are served with `vLLM`'s PagedAttention implementation (Kwon et al., 2023) in `bf16` precision, except for releases that require MXFP4 quantization, such as `gpt-oss`. The evaluation records log-probabilities for both prompt tokens and generated tokens, and sets `max_tokens` to 32,768. All runs are performed on clusters equipped with $8\times$ NVIDIA H200 GPUs with 141 GB of memory per GPU.

J.8 Token Statistics

The 20-configuration TTS response banks cover four benchmarks, 30 questions per benchmark, and 80 stochastic trials per model–question pair, for 192,000 inference runs. They generate 2.96B tokens in total. Prompt tokens account for 37M tokens (1.2%), and completion tokens account for 2.93B tokens (98.8%), giving an average of 15,434 tokens per query. Table 36 reports task-level completion-token counts for these banks and the AIME'26 temperature sweep.

Table 36: Task-level completion-token counts. AIME'24, AIME'25, HMMT'25, and BrUMO'25 are aggregated over 20 model configurations, 80 stochastic trials, and 30 questions per task. The AIME'26 row is the Qwen3.5-4B temperature-sweep response bank.

Task	**Completion tokens (M)**
AIME'24	680.0
AIME'25	728.3
HMMT'25	851.2
BrUMO'25	666.9
AIME'26 temperature sweep	272.5
Total	**3,198.9**

K Failure Modes of Repeated-Sampling Metrics: pass@k

Repeated-sampling evaluation starts from the finite response bank introduced in Section 2.1, then reports one or more functionals of that bank.

The shared finite-bank notation is

$$R \in \{0,1\}^{M\times N}, \qquad \gamma_i = \sum_{j=1}^{N} R_{ij}, \qquad h_c = \#\{i : \gamma_i = c\}, \quad c = 0,\dots,N,$$

where rows are questions, columns are sampled attempts, and $R_{ij} = 1$ denotes a correct attempt. For a budget $k \le N$ and threshold $t \in \{1,\dots,k\}$, the finite-bank threshold profile draws k responses without replacement from each row:

$$\widehat{S}^{\mathrm{fs}}_{k,t}(R) = \frac{1}{M}\sum_{i=1}^{M}\sum_{r=t}^{k} \frac{\binom{\gamma_i}{r}\binom{N-\gamma_i}{k-r}}{\binom{N}{k}}. \tag{44}$$

Invalid binomial coefficients are interpreted as zero. The usual finite-bank pass@k estimator is the $t = 1$ coordinate of (44) (Chen et al., 2021):

$$\mathrm{pass}_i\,@k(R) = \psi_{k,N}(\gamma_i), \qquad \psi_{k,N}(c) = 1 - \frac{\binom{N-c}{k}}{\binom{N}{k}}, \tag{45}$$

$$\mathrm{pass@}k(R) = \frac{1}{M}\sum_{i=1}^{M}\left(1 - \frac{\binom{N-\gamma_i}{k}}{\binom{N}{k}}\right) = \frac{1}{M}\sum_{c=0}^{N} h_c\psi_{k,N}(c). \tag{46}$$

The corresponding latent-process notation writes p_i for the probability that an independent attempt on question i is correct and $X_{i,k} \mid p_i \sim \mathrm{Binomial}(k, p_i)$. The population threshold coordinate is

$$q_{k,t}(p_i) = \Pr\{X_{i,k} \ge t \mid p_i\}, \qquad S_{k,t} = M^{-1}\sum_{i=1}^{M} q_{k,t}(p_i).$$

Thus $\mathrm{pass@}k$ is one low-threshold coordinate of a larger discovery–stability shape, while average accuracy, coverage, stability, and intermediate repeatability are different functionals of the same response-count distribution.

The distinction between discovery and stability is visible even when average accuracy is fixed. Fix an integer number of correct cells $C \in \{0, \dots, MN\}$, so $\bar{a}_R = C/(MN)$. Since the finite-bank shape depends on a row only through its count γ_i, it is enough to compare row-count vectors with the same sum C. A stability-concentrated bank packs successes into as few question rows as possible: if $C = q_{\mathrm{s}}N + r_{\mathrm{s}}$ with $0 \le r_{\mathrm{s}} < N$, it has q_{s} full rows and, when $r_{\mathrm{s}} > 0$, one partial row. A discovery-concentrated bank spreads the same C successes as evenly as the integer constraints allow: if $C = q_{\mathrm{d}}M + r_{\mathrm{d}}$ with $0 \le r_{\mathrm{d}} < M$, it has r_{d} rows with $q_{\mathrm{d}} + 1$ successes and $M - r_{\mathrm{d}}$ rows with q_{d} successes. Using $\mathbf{1}_m\mathbf{0}_{N-m}$ for a row with m ones followed by $N - m$ zeros, representative arrangements are

$$R_{\mathrm{stab}}(C) = \begin{bmatrix} \mathbf{1}_N \\ \vdots \\ \mathbf{1}_N \\ \mathbf{1}_{r_{\mathrm{s}}}\mathbf{0}_{N-r_{\mathrm{s}}} \\ \mathbf{0}_N \\ \vdots \\ \mathbf{0}_N \end{bmatrix} \quad R_{\mathrm{disc}}(C) = \begin{bmatrix} \mathbf{1}_{q_{\mathrm{d}}+1}\mathbf{0}_{N-q_{\mathrm{d}}-1} \\ \vdots \\ \mathbf{1}_{q_{\mathrm{d}}+1}\mathbf{0}_{N-q_{\mathrm{d}}-1} \\ \mathbf{1}_{q_{\mathrm{d}}}\mathbf{0}_{N-q_{\mathrm{d}}} \\ \vdots \\ \mathbf{1}_{q_{\mathrm{d}}}\mathbf{0}_{N-q_{\mathrm{d}}} \end{bmatrix}.$$

The partial row in $R_{\mathrm{stab}}(C)$ is omitted when $r_{\mathrm{s}} = 0$, and the upper block in $R_{\mathrm{disc}}(C)$ contains r_{d} rows. Both banks have average accuracy $\bar{a}_R$, but their threshold profiles encode different subsampling probabilities before any latent-probability model is imposed.

The first three failure modes assume that R is an exact binary correctness matrix, attempts are usable as independent samples, incorrect attempts carry no separate cost, and the finite bank is summarized without statistical uncertainty. Under these assumptions, the failure is scalar projection: a single $\mathrm{pass@}k$ value can change rankings across budgets, converge to coverage while average accuracy remains separate, or identify a large equivalence class of response-bank shapes. The later failure modes relax the auxiliary assumptions. Verifier errors can make the observed bank a matrix of acceptances; dependence can miscalibrate the budget axis; harmful or costly attempts require outcome categories beyond binary success; and finite-bank estimates need uncertainty quantification before they are used for ranking.

K.1 Ranking Instability of $\mathrm{pass@}k$

A small finite-bank example makes the ranking instability visible. Let rows denote questions and columns denote sampled responses, with 1 indicating a correct response:

$$R_1 = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&0&0&0&0&0&0 \\ 1&1&1&1&1&1&1&1&1&1 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 1&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&1&1&0&0&0 \\ 1&1&1&1&1&1&1&1&0&0 \end{bmatrix}. \tag{47}$$

Finite-bank $\mathrm{pass@}k$ depends on these matrices through their row success counts: $(0, 2, 2, 4, 10)$ for R_1 and $(1, 1, 1, 6, 8)$ for R_2. Using the estimator in (46), the sign sequence of $\mathrm{pass@}k(R_1) - \mathrm{pass@}k(R_2)$ over $k = 1, \dots, 10$ is

$$+,\ -,\ +,\ +,\ +,\ -,\ -,\ -,\ -,\ -.$$

Thus R_1 has the larger $\mathrm{pass@}k$ value at $k = 1$ and $k = 3, 4, 5$, whereas R_2 has the larger value at $k = 2$ and $k = 6, \dots, 10$. The ordering changes at budgets $k = 2$, $k = 3$, and $k = 6$, as shown in Figure 26.

Changing the budget changes which projection of the row-count shape is observed. For two models with histograms $h^{(1)}$ and $h^{(2)}$ under the common finite-bank setup, define

$$\Delta_k = \mathrm{pass@}k(h^{(1)}) - \mathrm{pass@}k(h^{(2)}), \qquad k = 1, \dots, N.$$

Model comparison over budgets reduces to the sign sequence of $\Delta_1, \dots, \Delta_N$. A single sign change is weak instability; a *multiple reversal* has two or more strict sign changes after deleting zero entries.

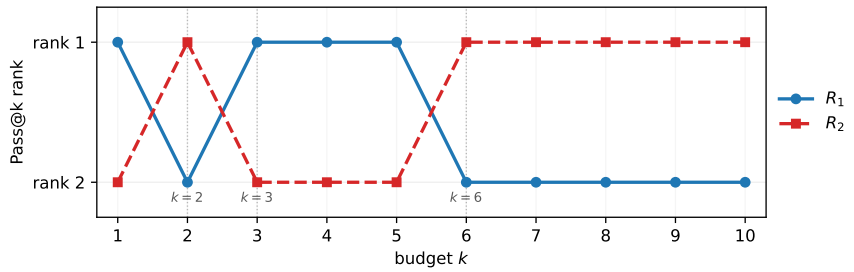

Figure 26: pass@k rank path for the response matrices in (47). Rank 1 denotes the larger pass@k value. The ordering changes at $k = 2$, $k = 3$, and $k = 6$.

The algebraic characterization below holds for arbitrary finite banks, and the simulations show that strict multiple reversals occupy non-negligible regions of profile space.

Definition 1 (Strict sign changes and multiple reversals)**.** *For a real sequence $x_1, \ldots, x_N$, let $V(x_1, \ldots, x_N)$ denote the number of sign changes after deleting zero entries. The pair $(h^{(1)}, h^{(2)})$ has $R = V(\Delta_1, \ldots, \Delta_N)$ budget-induced ranking reversals. The instability is called* multiple *when $R \geq 2$.*

K.1.1 Algebraic View of the Budget Path

Write

$$g_f = h_{N-f}, \qquad f = 0, \ldots, N,$$

for the corresponding failure-count histogram. For two models, define the signed normalized failure-profile difference

$$\sigma_f = \frac{g_f^{(2)} - g_f^{(1)}}{M} = \frac{h_{N-f}^{(2)} - h_{N-f}^{(1)}}{M}, \qquad \sum_{f=0}^{N} \sigma_f = 0. \tag{48}$$

Proposition 1 (pass@k is a normalized binomial transform)**.** *For every $k = 1, \ldots, N$,*

$$\Delta_k = \frac{1}{\binom{N}{k}} \sum_{f=0}^{N} \sigma_f \binom{f}{k}. \tag{49}$$

Equivalently, if

$$B_\sigma(z) = \sum_{f=0}^{N} \sigma_f (1+z)^f, \tag{50}$$

then

$$\Delta_k = \frac{[z^k] B_\sigma(z)}{\binom{N}{k}}. \tag{51}$$

Proof. For a row with $f = N - c$ failures, the failure probability at budget k is

$$1 - \text{pass}_i\,@k = \frac{\binom{f}{k}}{\binom{N}{k}}.$$

Therefore

$$\text{pass@}k(h^{(1)}) - \text{pass@}k(h^{(2)}) = \frac{1}{\binom{N}{k}} \sum_{f=0}^{N} \frac{g_f^{(2)} - g_f^{(1)}}{M} \binom{f}{k},$$

which proves (49). The coefficient identity follows from $[z^k](1+z)^f = \binom{f}{k}$. Q.E.D.

The proposition gives an exact algebraic characterization: pass@k rankings are the signs of the nonconstant coefficients of a shifted-basis polynomial. Opposite strict endpoint orderings certify at least one reversal, while the full coefficient sequence of B_σ can oscillate many times.

The same quantity can be written in terms of factorial moments of the row success count. Let

$$m_r(h) = \frac{1}{M} \sum_{c=0}^{N} h_c \binom{c}{r}.$$

By inclusion-exclusion,

$$\text{pass@}k(h) = \sum_{r=1}^{k} (-1)^{r+1} \frac{\binom{k}{r}}{\binom{N}{r}} m_r(h). \tag{52}$$

Increasing k changes the statistic itself by introducing higher-order factorial moments with alternating signs. For example,

$$\text{pass@}1(h) = \frac{m_1(h)}{N}, \qquad \text{pass@}2(h) = \frac{2m_1(h)}{N} - \frac{m_2(h)}{\binom{N}{2}}. \tag{53}$$

The second term in $\text{pass@}2(h)$ penalizes within-question concentration of successes. A model can have higher average correctness but also higher concentration, making it win at $k = 1$ and lose at $k = 2$.

At the other end of the budget path,

$$\text{pass@}N(h) = 1 - \frac{h_0}{M}, \tag{54}$$

so the final budget reduces to observed coverage, $1 - h_0/M$. More generally, for $r = 0, \ldots, N-1$,

$$1 - \text{pass@}(N-r)(h) = \frac{1}{\binom{N}{r} M} \sum_{c=0}^{r} h_c \binom{N-c}{r-c}. \tag{55}$$

Budgets close to N depend on the low-success tail $c \le r$. The tail dependence explains why rankings can change again after an intermediate-budget comparison: early budgets compare mean and concentration, while late budgets compare the mass of hard or unsolved questions.

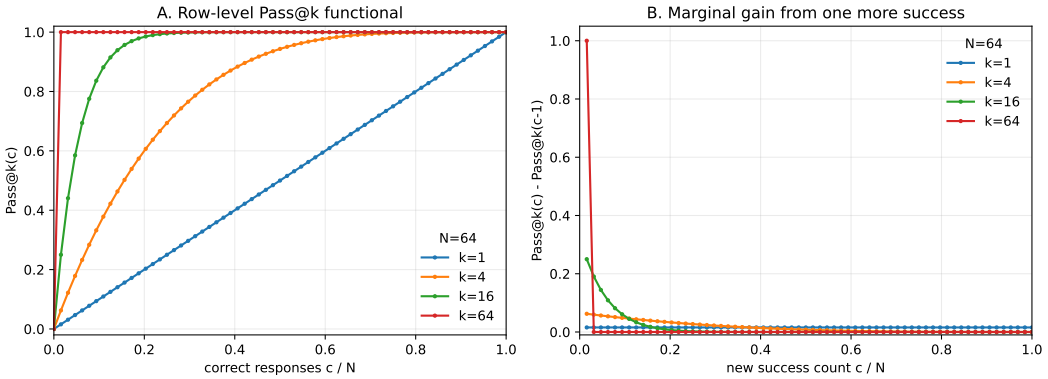

Figure 27: Deterministic budget functionals for a finite response bank with $N = 64$, visualized at the row level as the map from a row count c to $\text{pass}_i\,@k$. At $k = 1$ the functional is linear in the number of correct samples, while large budgets saturate after the first few successes and behave more like coverage over questions. Panel B gives the corresponding marginal gains: high-budget $\text{pass@}k$ assigns most of its incremental weight to moving a question from zero to a small positive number of correct responses.

K.1.2 Stability & Reversal Conditions

The simplest stability condition is stochastic dominance of the row-count distribution.

Proposition 2 (First-order dominance implies ranking stability)**.** *If the row-count distribution of $h^{(1)}$ first-order stochastically dominates that of $h^{(2)}$, i.e.,*

$$\frac{1}{M} \sum_{c=t}^{N} h_c^{(1)} \ge \frac{1}{M} \sum_{c=t}^{N} h_c^{(2)} \qquad \textit{for all } t = 1, \ldots, N,$$

then $\text{pass@}k(h^{(1)}) \ge \text{pass@}k(h^{(2)})$ *for every* $k = 1, \ldots, N$.

Proof. For each fixed k, the row functional $c \mapsto 1 - \binom{N-c}{k}/\binom{N}{k}$ is nondecreasing in c. Taking expectations under a first-order stochastic dominance relation preserves the inequality. Q.E.D.

A reversal entails crossing success-count profiles; a uniform per-question advantage gives the same ranking at every budget.

A sharper necessary condition comes from total positivity. The matrix $(\binom{f}{k})_{0\le f,k\le N}$ is the Pascal matrix, whose minors are nonnegative (total nonnegativity) (Pinkus, 2010). Positive scaling by $\binom{N}{k}^{-1}$ preserves signs, and the resulting transform cannot increase the number of strict sign changes.

Proposition 3 (A necessary condition for multiple reversals)**.** *Let $V(\sigma_0, \dots, \sigma_N)$ be the number of sign changes in the signed failure-profile difference after deleting zeros. Then*

$$V(\Delta_1, \dots, \Delta_N) \le V(\sigma_0, \dots, \sigma_N). \tag{56}$$

Equivalently, the number of budget-induced ranking reversals is at most the number of sign changes in the signed profile difference $h^{(1)} - h^{(2)}$ over success counts.

Proof sketch. Equation (49) applies a positively scaled Pascal transform to the signed vector σ. The Pascal kernel is totally nonnegative, and totally nonnegative transforms are variation-diminishing: the number of sign changes weakly decreases. Reversing the order from failures f to successes $c = N - f$ and multiplying by -1 preserves the sign-change count. Q.E.D.

The variation bound gives a diagnostic check. If the profile difference crosses only once, the $\mathrm{pass@}k$ path can exhibit at most one strict reversal. Multiple reversals require the models to exchange advantage across at least three separated correctness regimes.

When profiles are allowed to vary, the metric permits arbitrary sign patterns. The observed phenomenon is an algebraic feature of $\mathrm{pass@}k$.

Theorem 1 (Universal construction)**.** *Fix $N \ge 1$. For every sign vector $s = (s_1, \dots, s_N) \in \{-1, +1\}^N$, there exist a finite number of questions M and two integer histograms $h^{(1)}, h^{(2)}$ on $\{0, \dots, N\}$ such that*

$$\operatorname{sgn}\big(\mathrm{pass@}k(h^{(1)}) - \mathrm{pass@}k(h^{(2)})\big) = s_k \qquad \textit{for every } k = 1, \dots, N. \tag{57}$$

In particular, choosing alternating signs gives $N - 1$ strict ranking reversals.

Proof. Define a polynomial

$$B(z) = \sum_{k=1}^{N} s_k \binom{N}{k} z^k.$$

Since $(1+z)^0, \dots, (1+z)^N$ form a basis for polynomials of degree at most N, write

$$B(z) = \sum_{f=0}^{N} a_f (1+z)^f.$$

The coefficients a_f are integers because $z^k = ((1+z)-1)^k$ has integer coefficients in the shifted basis. Moreover, $\sum_f a_f = B(0) = 0$.

Construct two failure-count histograms by separating the positive and negative parts:

$$g_f^{(2)} = (a_f)_+, \qquad g_f^{(1)} = (-a_f)_+.$$

Because $\sum_f a_f = 0$, the two histograms have the same total mass M. Convert them to success-count histograms by $h_c^{(a)} = g_{N-c}^{(a)}$. Then $\sigma_f = (g_f^{(2)} - g_f^{(1)})/M = a_f/M$. By Proposition 1,

$$\mathrm{pass@}k(h^{(1)}) - \mathrm{pass@}k(h^{(2)}) = \frac{[z^k]B(z)}{M\binom{N}{k}} = \frac{s_k}{M},$$

which proves the claim. Q.E.D.

The construction in Theorem 1 is worst-case. Strict multiple reversals still occupy open neighborhoods of the profile simplex.

Proposition 4 (Strict sign patterns are robust)**.** *Suppose a pair of normalized histograms has margin*

$$\eta = \min_{1 \le k \le N} |\Delta_k| > 0.$$

If each normalized histogram is perturbed in ℓ_1 distance by at most ε and $2\varepsilon < \eta$, then every sign $\operatorname{sgn}(\Delta_k)$ is preserved. Hence every strict multiple reversal has an open neighborhood of profile pairs with the same reversal pattern.

Proof. For every k, the row-level $\mathrm{pass}_i\,@k$ value lies in $[0, 1]$. Changing one normalized histogram by ℓ_1 distance ε changes its dataset-level $\mathrm{pass@}k$ value by at most ε. Perturbing both models changes Δ_k by at most 2ε. If $2\varepsilon < \eta$, no sign can cross zero. Q.E.D.

K.1.3 Numerical Checks

The binomial-transform representation characterizes the entire budget path, Theorem 1 shows that the metric can realize any strict sign pattern, and Proposition 4 shows that strict patterns occupy open neighborhoods. The numerical checks diagnose these algebraic results.

Broad random-profile simulation. For a broad profile-space check, pairs of normalized success-count histograms were sampled independently from

$$p^{(1)}, p^{(2)} \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_{N+1})$$

on the simplex over success counts $\{0, \ldots, N\}$. For each pair, $\Delta_1, \ldots, \Delta_N$ was evaluated from (46), and $R = V(\Delta_1, \ldots, \Delta_N)$ was recorded. Table 37 and Figure 28 report 20,000 independent pairs per cell for $N \in \{8, 16, 32, 64, 128\}$ and $\alpha \in \{0.1, 0.3, 1, 3\}$. These estimates measure reversal probabilities under the specified Dirichlet distributions. The Monte Carlo standard error of each displayed probability is at most $\sqrt{0.25/20000} \approx 0.0035$.

Table 37: Broad random-profile simulation. Entries are estimated $\Pr[R \ge 2]$ for independent Dirichlet-distributed normalized profiles on $\{0, \ldots, N\}$; each cell uses 20,000 pairs.

N	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 1$	$\alpha = 3$
8	0.079	0.055	0.040	0.040
16	0.129	0.088	0.073	0.067
32	0.172	0.126	0.110	0.101
64	0.222	0.173	0.149	0.138
128	0.269	0.221	0.193	0.186

Across all displayed settings, multiple reversals appear with non-negligible probability. For example, at $N = 128$ the estimated probability of $R \ge 2$ ranges from 0.186 to 0.269 across the four concentration parameters. The trend with N is descriptive for this Dirichlet family and supports the interpretation that multiple reversals extend beyond small hand-constructed profile pairs.

Small exact finite-integer sanity check. Exact enumeration provides a finite-integer check on the reversal counts. Enumerating all sorted profiles $0 \le \nu_1 \le \cdots \le \nu_5 \le 8$ gives $\binom{13}{5} = 1287$ profiles and $827{,}541$ unordered pairs. Table 38 reports the sign-change counts in this exhaustive scan. The exact scan remains small because exhaustive enumeration is combinatorial in both M and N.

Practical interpretation. The diagnostic implication is to inspect the budget path. When comparing two models, the sequence $\Delta_1, \ldots, \Delta_N$ should be checked for $R \ge 2$ after ties are removed. Multiple sign changes indicate different distributions of row success counts, making a one-budget ranking an unreliable summary of test-time scaling behavior.

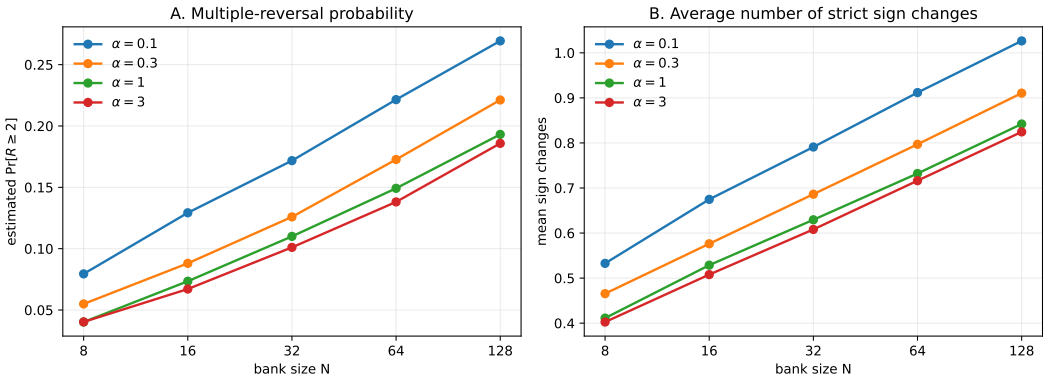

Figure 28: Broad random-profile simulation over the success-count simplex. Panel A reports the estimated probability of at least two strict pass@k ranking reversals; Panel B reports the mean number of strict sign changes. The increasing trend with N is consistent with the algebraic sign-pattern construction: longer budget paths give the binomial-transform coefficient sequence more opportunities to oscillate.

Table 38: Exact enumeration of sorted profiles for $M = 5, N = 8$. A multiple reversal means at least two strict sign changes in $\Delta_1, \ldots, \Delta_N$ after deleting ties.

Number of sign changes	Unordered profile pairs
0	661,388
1	163,076
2	3,077
At least one	166,153
At least two	3,077

K.2 Large-Budget Non-Convergence of pass@k to Average Accuracy

Large-budget discovery and average single-sample accuracy are different functionals of the same per-question success distribution. Augment the common finite-bank setup with conditionally independent attempts

$$R_{i,j} \mid p_i \sim \text{Bernoulli}(p_i), \qquad j = 1, \ldots, N,$$

where $p_i \in [0, 1]$ is the one-sample correctness probability for question i. Across questions, assume $p_1, p_2, \ldots$ are i.i.d. draws from a difficulty distribution F. The infinite-bank average-accuracy target is

$$A \equiv \mathbb{E}_F[p], \tag{58}$$

with finite-bank empirical version

$$A_{M,N} = \frac{1}{M} \sum_{i=1}^{M} \frac{\gamma_i}{N} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} R_{i,j}. \tag{59}$$

Equation (46) gives

$$\text{pass@1}(R) = A_{M,N}. \tag{60}$$

The target changes for every $k > 1$: pass@k measures at-least-one-success discovery, while expected correctness of a randomly drawn response remains the $k = 1$ target.

K.2.1 What the Large-Budget Regimes Converge to

For a fixed question with single-sample success probability p_i, define the population question-level discovery functional

$$\text{pass}_i\,@k(p_i) = 1 - (1 - p_i)^k, \qquad k = 1, 2, \ldots, \tag{61}$$

and the infinite-budget question-level functional

$$\mathrm{pass}_i\,@\infty(p_i) = \lim_{k\to\infty} \mathrm{pass}_i\,@k(p_i) = \mathbf{1}\{p_i > 0\}. \tag{62}$$

The corresponding population $\mathrm{pass}@k$ and coverage targets are

$$\mathrm{pass}@k = \mathbb{E}_F[\mathrm{pass}_i\,@k(p)], \qquad \mathrm{pass}@\infty = \mathbb{E}_F[\mathrm{pass}_i\,@\infty(p)] = \Pr_F(p > 0). \tag{63}$$

For the almost-sure limits, let $Z_i = (p_i, R_{i,1}, R_{i,2}, \ldots)$ be i.i.d. infinite rows, with responses conditionally i.i.d. Bernoulli(p_i) within each row. Every response bank is the leading $M \times N$ submatrix of this same array.

Proposition 5 (Asymptotic targets of finite-bank $\mathrm{pass}@k$)**.** *Let K_N be a fixed deterministic integer budget path with $1 \le K_N \le N$.*

1. *For fixed N and fixed $k \le N$, as $M \to \infty$,*

$$A_{M,N} \to A, \qquad \mathrm{pass}@k(R) \to \mathrm{pass}@k$$

almost surely.

2. *For fixed M, as $N \to \infty$,*

$$A_{M,N} \to \frac{1}{M}\sum_{i=1}^{M} p_i \qquad \textit{almost surely.}$$

If $K_N \to k_0 < \infty$, then

$$\mathrm{pass}@K_N(R) \to \frac{1}{M}\sum_{i=1}^{M} \mathrm{pass}_i\,@k_0(p_i) \qquad \textit{almost surely,}$$

whereas if $K_N \to \infty$, then

$$\mathrm{pass}@K_N(R) \to \frac{1}{M}\sum_{i=1}^{M} \mathrm{pass}_i\,@\infty(p_i) \qquad \textit{almost surely.}$$

3. *If $M, N \to \infty$ and $K_N \to k_0 < \infty$, then*

$$\mathrm{pass}@K_N(R) \to \mathrm{pass}@k_0.$$

If $M, N \to \infty$ and $K_N \to \infty$, then

$$\mathrm{pass}@K_N(R) \to \mathrm{pass}@\infty.$$

The analogous average-accuracy limit is always $A_{M,N} \to A$ whenever $M \to \infty$.

All convergences in the third item hold in probability and, under the nested construction above, almost surely. The joint limits require no relative growth-rate condition on M and N; the average-accuracy limit holds for any sequence $N = N(M) \ge 1$ as $M \to \infty$.

Proof. First, conditional on $p_i = p$, $\gamma_i \sim \mathrm{Binomial}(N, p)$ and

$$\mathbb{E}\left[\frac{\gamma_i}{N} \mid p_i = p\right] = p.$$

For $\mathrm{pass}@k$, select k indices uniformly without replacement from the N sampled responses. Conditional on the finite bank, the probability that all selected responses are failures is $\binom{N-\gamma_i}{k}/\binom{N}{k}$. Unconditionally, the selected responses are k independent model samples, so

$$\mathbb{E}\left[1 - \frac{\binom{N-\gamma_i}{k}}{\binom{N}{k}} \mid p_i = p\right] = 1 - (1-p)^k = \mathrm{pass}_i\,@k(p).$$

For fixed N, k, the rows are i.i.d. bounded random variables, so the first claim follows from the strong law of large numbers (Billingsley, 1995).

For fixed M and $N \to \infty$, the strong law gives $\gamma_i/N \to p_i$ for every fixed row. If $K_N \to k_0 < \infty$, continuity of the finite product representation of the hypergeometric failure probability gives convergence to $\text{pass}_i\,@k_0(p_i)$. If $K_N \to \infty$ and $p_i = 0$, then $\gamma_i = 0$ for all N and the row pass@K_N value is zero. If $p_i > 0$, then eventually $\gamma_i/N \geq p_i/2$. The failure probability is zero when $K_N > N - \gamma_i$; otherwise,

$$\frac{\binom{N-\gamma_i}{K_N}}{\binom{N}{K_N}} = \prod_{\ell=0}^{K_N-1} \frac{N-\gamma_i-\ell}{N-\ell} \leq \left(1 - \frac{p_i}{2}\right)^{K_N} \to 0.$$

This proves the fixed-M limits. For the joint limits, independence across questions and bounded row statistics give variance at most $1/(4M)$, which vanishes as $M \to \infty$. Chebyshev's inequality then gives convergence to the expectation. Dominated convergence gives $\mathbb{E}_F[\text{pass}_i\,@K_N(p)] \to \text{pass@}k_0$ or $\mathbb{E}_F[\text{pass}_i\,@K_N(p)] \to \text{pass@}\infty$ (Billingsley, 1995).

Under the nested construction, write $f_N(Z_i) \in [0,1]$ for the row statistic. The fixed-row limits give $f_N(Z_i) \to f(Z_i)$ almost surely, where $f(Z_i)$ is $1-(1-p_i)^{k_0}$, $\mathbf{1}\{p_i > 0\}$, or p_i for the respective cases. For each integer $L \geq 1$, define

$$g_L(Z_i) = \sup_{N \geq L} |f_N(Z_i) - f(Z_i)|.$$

Then $0 \leq g_L \leq 1$ and $g_L \downarrow 0$ almost surely, so $\mathbb{E}[g_L] \to 0$ by dominated convergence. Applying the strong law to f and to each of the countably many g_L gives, on one event of probability one,

$$\sup_{N \geq L} \left| \frac{1}{M}\sum_{i=1}^{M} f_N(Z_i) - \frac{1}{M}\sum_{i=1}^{M} f(Z_i) \right| \leq \frac{1}{M}\sum_{i=1}^{M} g_L(Z_i) \longrightarrow \mathbb{E}[g_L].$$

Consequently, for each L,

$$\limsup_{M,N\to\infty} \left| \frac{1}{M}\sum_{i=1}^{M} f_N(Z_i) - \mathbb{E}[f] \right| \leq \mathbb{E}[g_L] \qquad \text{almost surely.}$$

Letting $L \to \infty$ proves the joint almost-sure limits. For average accuracy, each of the finitely many $N < L$ row means also satisfies the strong law with expectation A. Combining these limits with the uniform bound for $N \geq L$ proves convergence for any $N = N(M) \geq 1$. Q.E.D.

Table 39: Asymptotic targets. Here $p \sim F$ is the random per-question single-sample correctness probability, $A = \mathbb{E}_F[p]$, $\text{pass@}\infty = \Pr_F(p > 0)$, and $\text{pass}_i\,@k(p) = 1-(1-p)^k$. The only path that recovers average accuracy from pass@k is $k = 1$ or the degenerate case $p \in \{0,1\}$ F-almost surely.

Regime	Budget path	Average-accuracy limit	pass@k limit
$M \to \infty$, fixed N, k	finite per-question budget	A	pass@k
$N \to \infty$, fixed M	fixed k	$M^{-1}\sum_i p_i$	$M^{-1}\sum_i \text{pass}_i\,@k(p_i)$
$N \to \infty$, fixed M	$k_N \to \infty$	$M^{-1}\sum_i p_i$	$M^{-1}\sum_i \mathbf{1}\{p_i > 0\}$
$M, N \to \infty$	fixed k	A	pass@k
$M, N \to \infty$	$k_N \to \infty$	A	pass@∞

The limit $M \to \infty$ removes question-level Monte Carlo noise for the budget already chosen. The limit $N \to \infty$ removes finite-bank noise and permits larger k. Once $k_N \to \infty$, the question-level target is $\text{pass}_i\,@\infty(p_i) = \mathbf{1}\{p_i > 0\}$, so the $N \to \infty$ and $(M, N) \to \infty$ regimes collapse to empirical or population coverage. Increasing M converts empirical coverage into population coverage; average accuracy remains a separate target.

K.2.2 pass@k Exceeds Average Accuracy for Nontrivial Success Probabilities

Theorem 2 (Non-convergence to average accuracy)**.** *For every integer $k \geq 1$,*

$$\text{pass@}k - A = \mathbb{E}_F\left[(1-p)\left\{1-(1-p)^{k-1}\right\}\right] \geq 0. \tag{64}$$

For $k > 1$, equality holds if and only if $p \in \{0, 1\}$ F-almost surely. Moreover,

$$\text{pass@}\infty - A = \mathbb{E}_F\left[\mathbf{1}\{p > 0\}(1-p)\right] \geq 0, \tag{65}$$

with equality if and only if $p \in \{0, 1\}$ F-almost surely. Therefore, for any question distribution with positive mass on $0 < p < 1$, large-budget pass@k *has a coverage limit that differs from infinite-bank average accuracy.*

Proof. For fixed $p \in [0, 1]$,

$$\text{pass}_i\,@k(p) - p = 1 - (1-p)^k - p = (1-p)\left\{1 - (1-p)^{k-1}\right\}.$$

This proves (64) after taking expectations. For $k > 1$, the integrand is strictly positive exactly when $0 < p < 1$, and zero at $p = 0$ and $p = 1$. Letting $k \to \infty$ and using bounded convergence gives (65). Q.E.D.

pass@1 equals average accuracy, but every $k > 1$ applies a concave saturation map to each question. As k grows, the map approaches a step function: any question with any nonzero probability of success receives full credit.

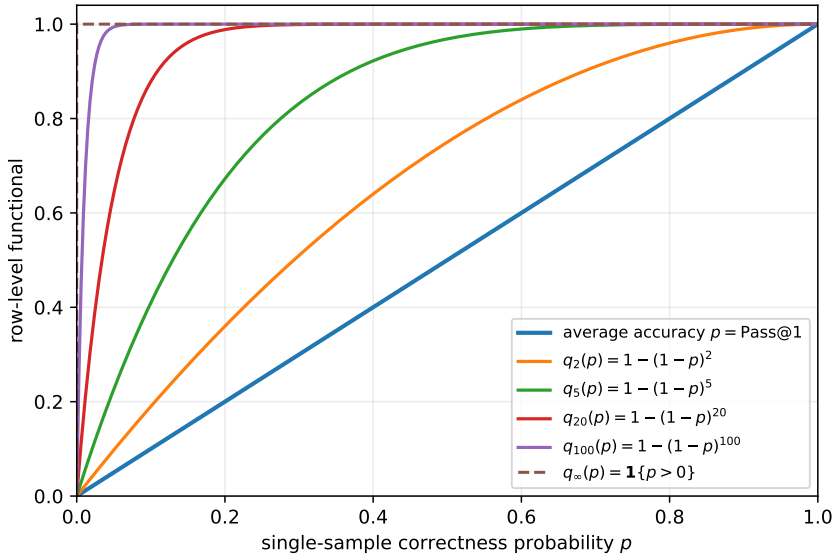

Figure 29: Population question-level functionals. Average accuracy is the identity map $p \mapsto p$, which equals pass@1. pass@k for $k > 1$ is $\text{pass}_i\,@k(p) = 1 - (1-p)^k$, an at-least-one-success transform. As $k \to \infty$, the functional becomes coverage, $\mathbf{1}\{p > 0\}$.

K.2.3 Explicit Counterexamples

The simplest counterexample is homogeneous. Suppose every question has the same single-sample correctness probability $p_i \equiv \varepsilon$, where $0 < \varepsilon < 1$. Then

$$A = \varepsilon, \qquad \text{pass@}k = 1 - (1-\varepsilon)^k, \qquad \text{pass@}\infty = 1. \tag{66}$$

Thus the large-budget pass@k limit is 1, while average accuracy is only ε. The gap $1 - \varepsilon$ can be made arbitrarily close to one by taking ε small.

The mismatch also reverses model rankings. Consider two population models:

$$\begin{aligned} &\text{Model A:} \quad p_A = 0.8 \text{ on half the questions, and } p_A = 0 \text{ otherwise,} \\ &\text{Model B:} \quad p_B = 0.1 \text{ on every question.} \end{aligned} \tag{67}$$

Average accuracy ranks A above B:

$$A_A = 0.5 \cdot 0.8 = 0.4, \qquad A_B = 0.1.$$

Table 40: Exact non-convergence examples. $A = \mathbb{E}_F[p]$ is the infinite-bank average accuracy and $\text{pass@}\infty = \Pr_F(p > 0)$ is the large-budget $\text{pass@}k$ target.

Model family	A	pass@2	pass@10	pass@100	pass@∞	pass@$\infty - A$
Homogeneous weak solver: $p_i \equiv 0.01$	0.0100	0.0199	0.0956	0.6340	1.0000	0.9900
Homogeneous moderate solver: $p_i \equiv 0.20$	0.2000	0.3600	0.8926	1.0000	1.0000	0.8000
Zero-inflated: $0.25\delta_0 + 0.75\text{Beta}(2, 8)$	0.1500	0.2591	0.5921	0.7454	0.7500	0.6000

But large-budget $\text{pass@}k$ ranks B above A:

$$\text{pass}^A\ @\infty = 0.5, \qquad \text{pass}^B\ @\infty = 1.$$

The finite-budget curves are

$$\text{pass}^A\ @k = 0.5\{1 - 0.2^k\}, \qquad \text{pass}^B\ @k = 1 - 0.9^k. \tag{68}$$

A wins at $k = 1, 2, \ldots, 6$, while B wins from $k = 7$ onward.

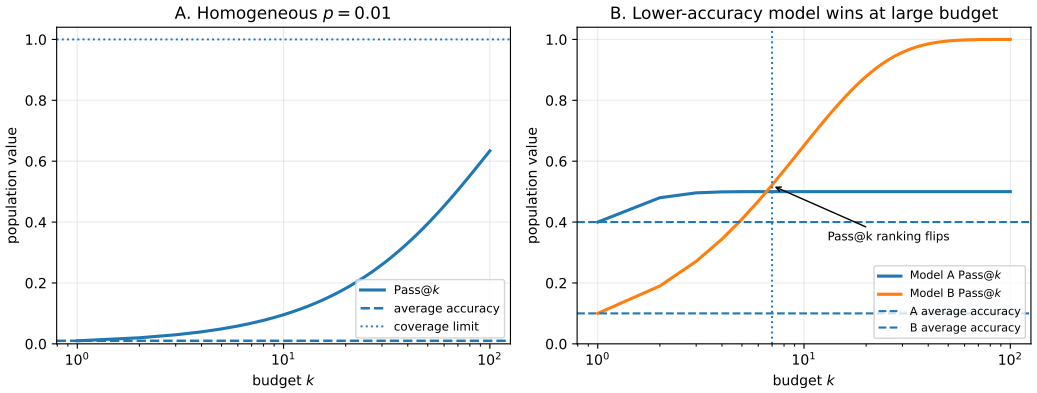

Figure 30: Two exact counterexamples. Panel A shows the homogeneous case $p_i \equiv 0.01$: average accuracy stays at 0.01, while $\text{pass@}k$ tends to coverage 1. Panel B shows a ranking inversion: Model A has higher average accuracy, but Model B has positive success probability on every question and wins under large-budget $\text{pass@}k$.

Table 41: A ranking counterexample. Model A has higher average accuracy: it solves half the tasks with single-sample probability 0.8 and never solves the other half, so $A_A = 0.4$. Model B has lower average accuracy $A_B = 0.1$ but positive probability on every task. Large-budget $\text{pass@}k$ ranks B above A.

Budget k	Model A $\text{pass@}k$	Model B $\text{pass@}k$	$\text{pass@}k$ winner
1	0.4000	0.1000	A
2	0.4800	0.1900	A
5	0.4998	0.4095	A
7	0.5000	0.5217	B
10	0.5000	0.6513	B
100	0.5000	1.0000	B
∞	0.5000	1.0000	B

K.2.4 Numerical Checks

A zero-inflated Beta family, a mixed discrete-continuous construction for unit-interval data with boundary mass (Ospina & Ferrari, 2010), models the random per-question success probability,

$$p \sim \pi\delta_0 + (1 - \pi)\,\text{Beta}(a, b). \tag{69}$$

This family separates two notions that $\mathrm{pass@}k$ conflates: the coverage fraction $1-\pi$ and the average accuracy on covered questions $a/(a+b)$. Its targets have closed forms:

$$A = (1-\pi)\frac{a}{a+b}, \qquad \mathrm{pass@}k = (1-\pi)\left\{1 - \frac{\mathrm{B}(a, b+k)}{\mathrm{B}(a,b)}\right\}, \qquad \mathrm{pass@}\infty = 1-\pi. \quad (70)$$

Finite-bank convergence. For $p \sim 0.25\delta_0 + 0.75\,\mathrm{Beta}(2,8)$, the average-accuracy target is $A = 0.15$ and the coverage target is $\mathrm{pass@}\infty = 0.75$. We simulated finite banks with $k = N$. Table 42 shows that increasing M reduces Monte Carlo error, while increasing $N = k$ pushes $\mathrm{pass@}N$ toward coverage.

Table 42: Finite-bank simulation for $p \sim 0.25\delta_0 + 0.75\mathrm{Beta}(2,8)$ with $k = N$. The average-accuracy target is $A = 0.15$ for every N, while the expected $\mathrm{pass@}N$ target moves toward coverage $\mathrm{pass@}\infty = 0.75$. Monte Carlo columns report means over 1,000 independent experiments with approximate 95% confidence intervals.

M	$N = k$	target A	target $\mathrm{pass@}N$	MC accuracy	MC $\mathrm{pass@}N$
50	5	0.1500	0.4533	0.1496±0.0017	0.4527±0.0043
50	20	0.1500	0.6835	0.1487±0.0014	0.6821±0.0041
50	100	0.1500	0.7454	0.1507±0.0012	0.7466±0.0038
500	5	0.1500	0.4533	0.1499±0.0006	0.4531±0.0014
500	20	0.1500	0.6835	0.1500±0.0004	0.6842±0.0013
500	100	0.1500	0.7454	0.1498±0.0004	0.7451±0.0012
5000	20	0.1500	0.6835	0.1499±0.0001	0.6832±0.0004
5000	200	0.1500	0.7488	0.1500±0.0001	0.7486±0.0004

Random model-pair check. For the random model-pair check, pairs of zero-inflated-Beta model families were sampled independently. For each model,

$$\pi \sim \mathrm{Unif}(0, 0.9), \qquad \mu = \frac{a}{a+b} \sim \mathrm{Unif}(0.01, 0.99), \qquad \tau = a + b \sim \mathrm{LogUnif}(2, 50),$$

were drawn independently, with $a = \mu\tau$ and $b = (1-\mu)\tau$. Exact A and $\mathrm{pass@}k$ values were computed from (70); no finite-bank sampling noise is involved in this ranking check. Table 43 reports the probability that the $\mathrm{pass@}k$ ranking disagrees with the average-accuracy ranking over 200,000 independent model pairs. The disagreement probability is zero at $k = 1$ by identity (60), then increases with the budget and reaches about 0.285 in the coverage limit.

Table 43: Random model-pair check. We sample two zero-inflated-Beta model families independently with $\pi \sim \mathrm{Unif}(0, 0.9)$, positive-task mean $\mu \sim \mathrm{Unif}(0.01, 0.99)$, and concentration $\tau \sim \mathrm{LogUnif}(2, 50)$. Entries estimate the probability that the $\mathrm{pass@}k$ ranking disagrees with the average-accuracy ranking over 200,000 independent model pairs.

Budget k	disagreement probability	approx. 95% half-width
1	0.000	0.000
2	0.048	0.001
5	0.120	0.001
10	0.168	0.002
20	0.202	0.002
50	0.230	0.002
100	0.243	0.002
∞	0.285	0.002

K.2.5 Practical Interpretation

The large-budget target of $\mathrm{pass@}k$ is coverage. If the intended estimand is expected correctness of a randomly drawn response, average accuracy or $\mathrm{pass@}1$ must be reported separately from large-budget discovery.

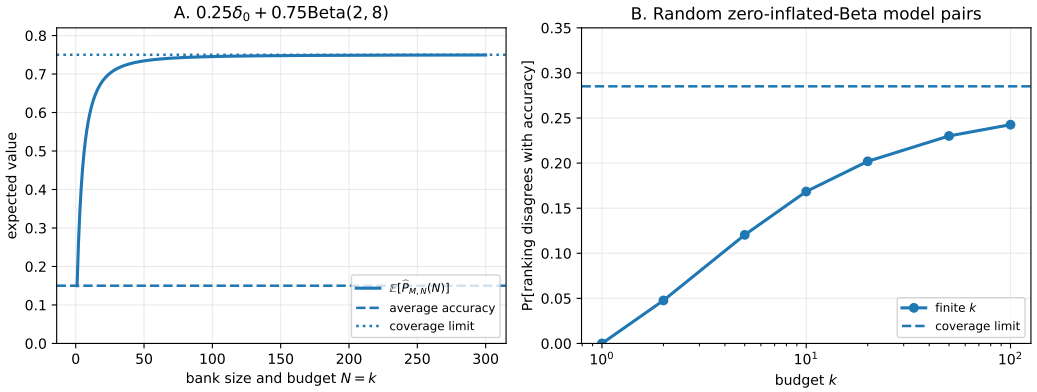

Figure 31: Numerical checks for the large-budget target. Panel A shows the exact expected pass@N value for $p \sim 0.25\delta_0 + 0.75\,\text{Beta}(2, 8)$ with $k = N$; it approaches the coverage value 0.75, while average accuracy is 0.15. Panel B shows the random model-pair disagreement probability from Table 43.

K.3 Structural Non-Identifiability of pass@k

Structural aliasing is deterministic loss of shape: distinct allocations of correct samples can receive identical scalar pass@k values. A concrete finite-bank alias makes the collapse visible. With two questions and four sampled responses per question,

$$R_A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad R_B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

The row counts are $(1, 1)$ for R_A and $(0, 3)$ for R_B. At budget $k = 2$, a row with one success contributes $1/2$, a row with no successes contributes 0, and a row with three successes contributes 1. Hence both banks have pass@$2 = 1/2$, even though R_A spreads one success across each question while R_B leaves one question unsolved and concentrates all successes on the other.

For a bank R, write $h(R)$ for its row-success histogram. For a set of reported budgets $K \subseteq \{1, \ldots, N\}$, write pass@$K(R) = (\text{pass@}k(R))_{k \in K}$, with pass@$k$ defined by (46).

Definition 2 (pass@K indistinguishability)**.** *Two response matrices R and R' are* pass@K-indistinguishable*, written $R \equiv_K R'$, if*

$$\text{pass@}k(R) = \text{pass@}k(R') \qquad \textit{for every } k \in K.$$

When $K = \{k\}$, we write $R \equiv_k R'$.

K.3.1 Two Sources of Structural Non-Identifiability

The first source is row-sum factorization: pass@k depends on the number of correct samples in a row, while the within-row support pattern is lost.

Proposition 6 (pass@k factors through the row-success histogram)**.** *For every $k = 1, \ldots, N$,* pass@$k(R)$ *depends on R only through $h(R)$. Hence any two matrices with the same row-success histogram are indistinguishable by every* pass@k *value, including the full* pass@$1{:}N$ *curve. More precisely, if an ordered row-count vector is $c_1, \ldots, c_M$, then*

$$\#\{R : \gamma_i(R) = c_i \textit{ for all } i\} = \prod_{i=1}^{M} \binom{N}{c_i}. \tag{71}$$

If only the histogram $h = (h_0, \ldots, h_N)$ is fixed, then the number of labeled response matrices with that histogram is

$$A(h) = \frac{M!}{\prod_{c=0}^{N} h_c!} \prod_{c=0}^{N} \binom{N}{c}^{h_c}. \tag{72}$$

Proof. Equation (46) uses row i only through the row count γ_i. For a fixed count c_i, there are $\binom{N}{c_i}$ binary rows with that count, giving (71). If only the histogram is fixed, the count classes can be assigned to labeled rows in $M!/\prod_c h_c!$ ways, giving (72). Q.E.D.

The second source is saturation of the row-level functional. At high budgets, once a row has enough successes, additional successes leave its row contribution unchanged.

Proposition 7 (Row-level resolution collapse)**.** *For* $c = 0, \dots, N-1$,

$$\psi_{k,N}(c+1) - \psi_{k,N}(c) = \frac{\binom{N-c-1}{k-1}}{\binom{N}{k}}. \tag{73}$$

Consequently, $\psi_{k,N}(c) = 1$ *for every* $c \geq N-k+1$*, and the number of distinct row-level values among* $c = 0, \dots, N$ *is*

$$L_{N,k} = N - k + 2. \tag{74}$$

Proof. Pascal's identity gives

$$\binom{N-c}{k} - \binom{N-c-1}{k} = \binom{N-c-1}{k-1}.$$

Substituting this identity into $\psi_{k,N}(c+1) - \psi_{k,N}(c)$ proves (73). The increment is positive for $c \leq N-k$ and zero for $c \geq N-k+1$, so the row functional reaches one at $c = N-k+1$ and remains there. Q.E.D.

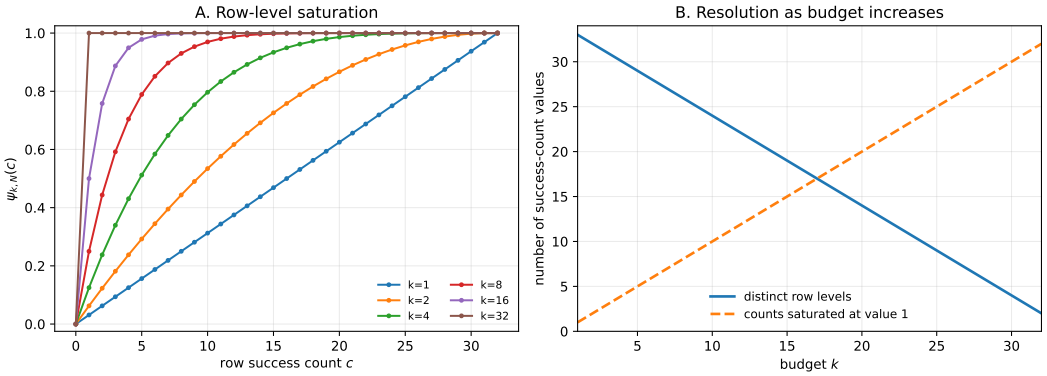

Figure 32: Row-level non-identifiability for $N = 32$. Panel A plots $c \mapsto \psi_{k,N}(c)$ for several budgets. As k increases, the map saturates earlier and collapses many success counts to the same value. Panel B shows the exact resolution count $L_{N,k} = N - k + 2$ and the number of success counts in the saturated class.

The endpoint and high-budget cases make the discarded information explicit. At $k = 1$,

$$\text{pass@1}(h) = \frac{1}{MN} \sum_{c=0}^{N} c h_c, \tag{75}$$

so pass@1 records only the total amount of success mass and forgets how that mass is distributed over questions. At the other endpoint,

$$\text{pass@}N(h) = 1 - \frac{h_0}{M}, \tag{76}$$

so pass@N is pure coverage: every nonzero row is merged, regardless of whether it has one success or N successes. More generally, writing $k = N - r$ gives

$$1 - \text{pass@}(N-r)(h) = \frac{1}{M\binom{N}{r}} \sum_{c=0}^{r} h_c \binom{N-c}{r-c}, \qquad r = 0, \dots, N-1. \tag{77}$$

Thus high-budget $\mathrm{pass@}k$ depends only on the low-success tail $h_0, \dots, h_r$ and is insensitive to how the remaining mass is distributed among $h_{r+1}, \dots, h_N$.

For intermediate budgets, $\mathrm{pass@}k$ is one alternating combination of factorial moments:

$$\mathrm{pass@}k(h) = \sum_{s=1}^{k} (-1)^{s+1} \frac{\binom{k}{s}}{\binom{N}{s}} \frac{1}{M} \sum_{c=0}^{N} h_c \binom{c}{s}. \tag{78}$$

Thus a fixed intermediate budget leaves the mean, coverage, tail mass, and concentration unidentified; it identifies only the particular linear functional in (78).

K.3.2 Equivalence Classes & Identifiability Rank

It is useful to clear denominators. Define

$$a_k(c) = \binom{N}{k} - \binom{N-c}{k}, \qquad T_k(h) = M\binom{N}{k}\,\mathrm{pass@}k(h) = \sum_{c=0}^{N} h_c a_k(c). \tag{79}$$

The scalar $T_k(h)$ is an integer in $\{0, \dots, M\binom{N}{k}\}$.

Proposition 8 (Exact kernel characterization)**.** *Fix $K \subseteq \{1, \dots, N\}$. Two histograms h and h' with total mass M are $\mathrm{pass@}K$-indistinguishable if and only if their signed difference $\delta_c = h_c - h'_c$ satisfies*

$$\sum_{c=0}^{N} \delta_c = 0, \qquad \sum_{c=0}^{N} \delta_c a_k(c) = 0 \quad \textit{for every } k \in K. \tag{80}$$

Equivalently, profile-level equivalence classes are affine lattice fibers of the linear map $h \mapsto (T_k(h))_{k\in K}$.

Proof. The first condition says that both histograms have the same number of rows. Equality of $\mathrm{pass@}k$ is equivalent to equality of the integer numerator $T_k(h)$ in (79), which is the second condition. The converse follows by reversing the same calculation. Q.E.D.

Let

$$\Delta_N = \left\{ \pi \in \mathbb{R}_+^{N+1} : \sum_{c=0}^{N} \pi_c = 1 \right\}$$

be the normalized profile simplex, and define $\Phi_K(\pi) = (\mathrm{pass@}k(\pi))_{k\in K}$. The exact continuous profile-level rank is as follows.

Theorem 3 (Identifiability rank of a budget set)**.** *Let $K \subseteq \{1, \dots, N\}$ and let $L = |K|$. At every interior point of Δ_N, the local fiber of Φ_K has affine dimension*

$$N - L. \tag{81}$$

Therefore any strict subset of budgets leaves a positive-dimensional family of indistinguishable success-count profiles. If $K = \{1, \dots, N\}$, then the full $\mathrm{pass@}1{:}N$ path identifies the row-success histogram exactly. The full path does not in general identify the underlying response matrix R.

Proof. Write the failure profile as $q_f = \pi_{N-f}$, so q_f is the mass on rows with f incorrect samples. For every $k = 1, \dots, N$,

$$1 - \mathrm{pass@}k(\pi) = \sum_{f=0}^{N} q_f \frac{\binom{f}{k}}{\binom{N}{k}}. \tag{82}$$

Together with the normalization row $k = 0$, these are binomial moments $\sum_f q_f \binom{f}{k}$. The matrix $(\binom{f}{k})_{0\le k, f\le N}$ is triangular with nonzero diagonal when rows and columns are ordered by increasing k and f, so its rows are linearly independent. Hence normalization plus the budgets in K impose $L+1$ independent affine constraints in $\mathbb{R}^{N+1}$. Inside the N-dimensional simplex, the local fiber dimension is $N - L$ at interior points. When $L = N$, the full set of binomial moments determines the failure profile q, and therefore π, uniquely. Proposition 6 shows that even this full curve leaves within-row supports unidentified. Q.E.D.

Table 44: Profile-level identifiability from a set K of reported budgets. The success-count profile simplex has dimension N and $L = |K|$.

Reported budgets	local profile-fiber dimension	Identifies response matrix?
one budget, $L = 1$	$N - 1$	no
proper subset, $1 < L < N$	$N - L$	no
full path, $L = N$	0 at profile level	no; within-row supports remain invisible

The continuous theorem has an exact integer counterpart, with one important qualification: for a fixed small M, a particular lattice fiber may contain only one feasible histogram. The non-identifiability is nevertheless unavoidable in finite banks once M is allowed to match an integer kernel direction.

Corollary 1 (Integer aliases for incomplete budget sets)**.** *If $|K| < N$, then there exist a finite M and two distinct histograms $h, h' \in \mathbb{Z}_+^{N+1}$ with $\sum_c h_c = \sum_c h'_c = M$ such that* $\mathrm{pass@}K(h) = \mathrm{pass@}K(h')$.

Proof. The linear system in (80) has integer coefficients and at most $|K| + 1 < N + 1$ independent rows. Its rational nullspace therefore contains a nonzero integer vector u with $\sum_c u_c = 0$ and $\sum_c u_c a_k(c) = 0$ for all $k \in K$. Writing $u = u_+ - u_-$ gives two nonnegative integer histograms $h = u_+$ and $h' = u_-$ with equal mass and equal $\mathrm{pass@}K$ values. Q.E.D.

K.3.3 Counting Bounds for Finite Banks

Let

$$\mathcal{H}_{M,N} = \{h \in \mathbb{Z}_+^{N+1} : \sum_c h_c = M\}, \qquad |\mathcal{H}_{M,N}| = \binom{M+N}{N},$$

and let $V_{M,N,k}$ be the number of distinct scalar $\mathrm{pass@}k$ values attained over $\mathcal{H}_{M,N}$.

Proposition 9 (Pigeonhole bounds for exact fibers)**.** *For every M, N, k,*

$$V_{M,N,k} \le \min\left\{ \binom{M+N-k+1}{N-k+1},\ M\binom{N}{k} + 1,\ \binom{M+N}{N} \right\}. \tag{83}$$

The average number of row-success histograms per attained scalar value is exactly

$$\frac{\binom{M+N}{N}}{V_{M,N,k}}, \tag{84}$$

and the average number of labeled response matrices per attained scalar value is exactly

$$\frac{2^{MN}}{V_{M,N,k}}. \tag{85}$$

For a budget set K, the number of possible score vectors is at most $\prod_{k \in K}(M\binom{N}{k} + 1)$.

Proof. By Proposition 7, there are only $L_{N,k} = N - k + 2$ distinct row-level values. A multiset of M rows over these levels has at most $\binom{M+L_{N,k}-1}{L_{N,k}-1}$ level histograms, giving the first bound in (83). The integer numerator $T_k(h)$ in (79) lies between 0 and $M\binom{N}{k}$, giving the second bound. The third bound is the number of histograms. The average-fiber identities follow by partitioning the histograms and the 2^{MN} labeled response matrices by their attained scalar values. The multi-budget bound applies the same integer range argument coordinatewise. Q.E.D.

The endpoint cases show two different forms of aliasing. For $k = 1$, the score depends only on the total number of correct samples, so $V_{M,N,1} = MN + 1$. For $k = N$, the score depends only on h_0, so $V_{M,N,N} = M + 1$.

K.3.4 Exact Enumeration Experiments

Exact finite-bank enumerations quantify the size of the induced equivalence classes. The main enumeration uses $M = N = 10$, for which there are $|\mathcal{H}_{10,10}| = 184{,}756$ row-success histograms. For every histogram and budget, the integer score $T_k(h)$ in (79) was computed. Matrix-level collision probabilities were computed exactly using the multiplicity formula

$$|\mathcal{E}_K(t)| = \sum_{h:T_K(h)=t} \frac{M!}{\prod_{c=0}^{N} h_c!} \prod_{c=0}^{N} \binom{N}{c}^{h_c}, \tag{86}$$

where $T_K(h) = (T_k(h))_{k\in K}$. No Monte Carlo estimates are used.

Table 45: Exact fixed-budget collapse for $M = 10$ questions and $N = 10$ samples. There are $|\mathcal{H}_{10,10}| = 184{,}756$ row-success histograms. The profile-pair column is the probability that two distinct uniformly chosen histograms have the same pass@k value. The matrix-pair column is the analogous exact probability for two distinct uniformly chosen binary response matrices, computed with row-pattern multiplicities.

k	row levels	values	mean class	max class	profile-pair (%)	matrix-pair (%)
1	11	101	1829.267	5,448	2.107	5.635
2	10	420	439.895	1,221	0.455	1.227
3	9	1,012	182.565	601	0.185	0.598
5	7	1,436	128.660	1,385	0.191	0.746
8	4	226	817.504	19,448	3.619	40.216
10	2	11	16796.000	92,378	34.126	98.074

Table 45 shows that a fixed budget has far fewer values than the number of profiles. Even the intermediate budget pass@5 has only 1,436 distinct values over 184,756 profiles, with a largest fiber containing 1,385 distinct row-success histograms. Endpoint coverage is much coarser: pass@10 has only 11 values. The matrix-pair collision probability at pass@10 is 98.074% because, under the uniform matrix measure, almost every row has at least one success and many matrices therefore share the same coverage score.

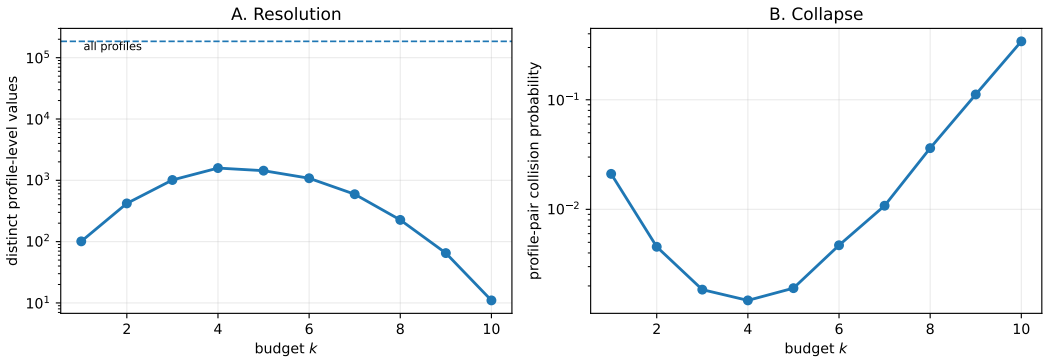

Figure 33: Exact profile-level resolution for $M = N = 10$. Panel A gives the number of distinct scalar pass@k values induced by all 184,756 row-success histograms; the dashed line is the number of histograms. Panel B gives the probability that two distinct uniformly chosen histograms collapse to the same pass@k value.

Adding budgets refines the equivalence relation. Table 46 reports exact results for several budget sets. The full pass@1:N curve has one profile per fiber, as predicted by Theorem 3. However, matrix-level collapse remains: for $M = N = 10$, the exact probability that two distinct uniformly chosen response matrices share the same full row-success histogram is 0.068%.

Fixed-budget fibers can still contain substantial variation. The within-class ranges are computed for three auxiliary quantities: pass@1, pass@N, and $\mathrm{Var}_h(\gamma/N)$. Table 47 shows that these quantities remain variable within fixed-budget classes. For pass@5, the mean within-class ranges are 0.192 for pass@1, 0.111 for pass@N, and 0.093 for normalized row-count variance.

Table 46: Adding budgets refines the profile equivalence relation for $M = 10, N = 10$. The full pass@1:N path identifies the row-success histogram, but not the underlying response matrix.

Budgets K	profile classes	mean class	max class	profile-pair (%)	matrix-pair (%)
$\{1\}$	101	1829.267	5,448	2.107	5.635
$\{5\}$	1,436	128.660	1,385	0.191	0.746
$\{10\}$	11	16796.000	92,378	34.126	98.074
$\{1, 10\}$	506	365.130	2,934	0.799	5.549
$\{1, 2, 10\}$	17,343	10.653	83	0.016	0.333
$\{1, 5, 10\}$	35,419	5.216	55	0.006	0.117
$\{1, \ldots, N\}$	184,756	1.000	1	0.000	0.068

Table 47: Auxiliary structure that remains variable inside fixed pass@k equivalence classes for $M = 10, N = 10$. Each entry is an exact range over histograms within an equivalence class, averaged over nonempty classes or maximized over classes.

	Mean within-class range			Maximum within-class range		
k	pass@1	pass@N	$\mathrm{Var}_h(\gamma/N)$	pass@1	pass@N	$\mathrm{Var}_h(\gamma/N)$
1	0.000	0.401	0.147	0.000	0.900	0.250
5	0.192	0.111	0.093	0.560	0.500	0.250
10	0.450	0.000	0.186	0.900	0.000	0.248

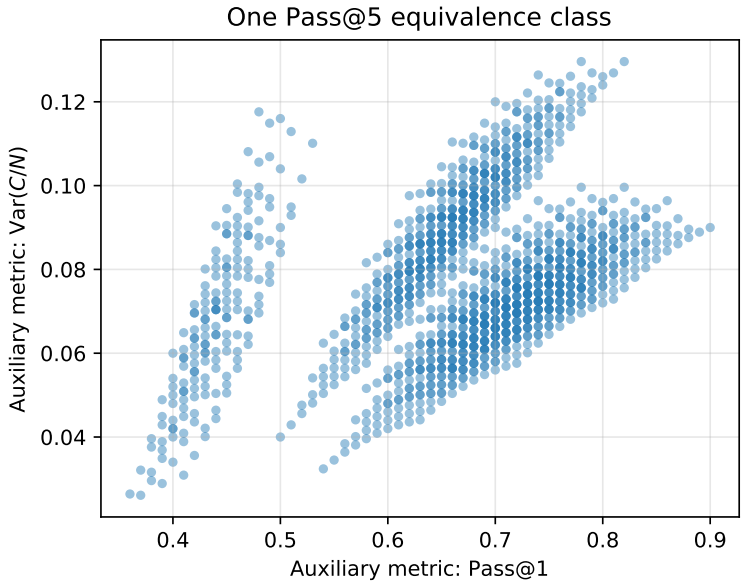

Figure 34: Auxiliary variation inside one pass@5 equivalence class for $M = N = 10$. The displayed class has pass@5 $= 0.90$ and contains 1,385 distinct row-success histograms. Within this single class, pass@1 ranges from 0.36 to 0.90, coverage pass@N ranges from 0.90 to 1.00, and $\mathrm{Var}_h(\gamma/N)$ ranges from 0.0261 to 0.1296.

A scaling scan repeats the exact profile-level enumeration for $M = N \in \{4, 6, 8, 10, 12\}$ at three representative budgets: 1, $\lceil N/2 \rceil$, and N. Table 48 and Figure 35 show two regimes. Endpoint coverage always has only $M + 1$ values. Intermediate budgets have higher resolution, but the number of row-success histograms grows much faster than the number of scalar scores; at $M = N = 12$ and $k = \lceil N/2 \rceil$, there are 2,704,156 histograms but only 5,628 distinct pass@k values.

Implications for test-time scaling evaluation. These results specify the equivalence classes behind the shape-level claim in Section 2.3. At the profile level, the full row-success histogram h/M, the

Table 48: Exact profile-level scaling scan with $M = N$. Endpoint coverage pass@N has only $M + 1$ values, while intermediate budgets have higher but still far-from-complete scalar resolution.

$N = M$	budget	profiles	values	profile-pair (%)
4	1	70	17	6.625
4	$\lceil N/2 \rceil$	70	21	5.135
4	N	70	5	34.617
6	1	924	37	4.328
6	$\lceil N/2 \rceil$	924	88	1.866
6	N	924	7	34.650
8	1	12,870	65	2.912
8	$\lceil N/2 \rceil$	12,870	345	0.761
8	N	12,870	9	34.338
10	1	184,756	101	2.107
10	$\lceil N/2 \rceil$	184,756	1,436	0.191
10	N	184,756	11	34.126
12	1	2,704,156	145	1.613
12	$\lceil N/2 \rceil$	2,704,156	5,628	0.089
12	N	2,704,156	13	33.987

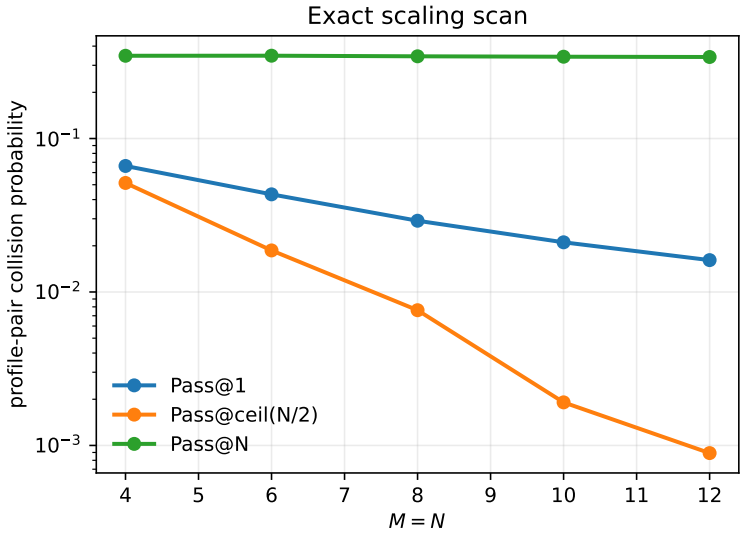

Figure 35: Exact profile-pair collision probabilities in the scaling scan $M = N \in \{4, 6, 8, 10, 12\}$. Endpoint pass@$N$ remains coarse because it records only the number of rows with zero successes. Intermediate budgets reduce random profile-pair collision probability, but their scalar fibers still contain many profiles.

pass@1:N curve, or a compact moment/tail suite such as $(\mathbb{E}_h[\gamma/N], \mathrm{Var}_h(\gamma/N), h_0/M)$ exposes distinctions that a one-budget scalar omits. At the matrix level, column-load dispersion and pairwise sample overlap diagnose whether successes are concentrated in a few samples or spread through the response bank. Even matching average accuracy and both discovery–stability endpoints leaves intermediate thresholds unresolved in the exact enumeration of Appendix D.1.

L Auxiliary Assumption Failures

The projection failures in Sections K.1, K.2, and K.3 assume that the observed response bank is an exact binary response matrix, that the sampling budget corresponds to independent usable attempts, that incorrect attempts have no independent cost, and that the finite bank is precise enough for ranking. These assumptions are often as important as the scalar functional itself. When any of them fails,

reporting only $\mathrm{pass@}k$ can misdescribe the relevant discovery–stability shape even before that shape is compressed to a scalar.

Table 49: Auxiliary failure modes for repeated-sampling evaluation. The scalar-projection failures in Appendix K assume exact binary labels, independent usable attempts, zero-cost incorrect attempts, and sufficiently precise finite-bank estimates; these modes arise when those auxiliary assumptions fail.

Mode	Hidden assumption	Failure diagnostic
Verifier/selector fragility	The correctness oracle is exact and a correct candidate can be selected from the batch.	False positives accumulate with k; reported accepted $\mathrm{pass@}k$ can approach one while the returned candidate has bounded precision.
Dependence/effective sample size	The k attempts behave like independent draws, or the finite-bank curve reveals enough about dependence.	Positive dependence makes nominal budget k equivalent to a much smaller independent budget.
Harmful-attempt blindness	Incorrect attempts have zero cost and only the existence of one correct attempt matters.	Two systems can have identical $\mathrm{pass@}k$ but arbitrarily different harmful-attempt probability or utility.
Finite-bank uncertainty	The point estimate is sufficiently precise for ranking.	The unbiased estimator has nontrivial U-statistic variance; near-saturated high-k scores can still induce false rankings.

L.1 Verifier & Selector Fragility of $\mathrm{pass@}k$

An observed response bank may record verifier acceptance as the correctness label. Unit tests for code, answer matching for mathematics, and LLM judges for open-ended reasoning are all measurement procedures. The finite-bank $\mathrm{pass@}k$ protocol measures the chance that a uniformly sampled k-subset contains a test-passing candidate (Chen et al., 2021). An exact verifier and a selector that returns a correct candidate whenever one exists are sufficient to equate this score with returned-candidate correctness. In practice, code benchmarks can have incomplete tests, and stronger test suites such as EvalPlus have shown that previously accepted programs may fail additional tests (Liu et al., 2023a). For math and reasoning evaluations, recent G-$\mathrm{Pass@}k_\tau$ work explicitly uses an LLM judge to compare model answers to references (Liu et al., 2025). Independent work on inference-time resampling with imperfect verifiers shows that false positives impose fundamental limits on repeated sampling (Stroebl et al., 2024).

Two gaps are distinct. When the reported label is verifier acceptance, false positives are amplified by the discovery transform. When selection from the batch is imperfect, returned-candidate correctness can be far below the reported $\mathrm{pass@}k$ value.

Setup. For a single question, let $Y_j \in \{0,1\}$ denote true correctness of candidate j and let $V_j \in \{0,1\}$ denote verifier acceptance. Assume candidates are i.i.d. and write

$$p = \Pr(Y=1), \qquad r = \Pr(V=1 \mid Y=1), \qquad f = \Pr(V=1 \mid Y=0). \tag{87}$$

Here r is the verifier true-positive rate and f is the false-positive rate. The per-sample acceptance probability is

$$a = \Pr(V=1) = rp + f(1-p). \tag{88}$$

The reported accepted $\mathrm{pass@}k$ is

$$\mathrm{pass}^{\mathrm{obs}}\,@k = \Pr\left(\max_{1\le j\le k} V_j = 1\right) = 1-(1-a)^k, \tag{89}$$

whereas the true discovery probability is

$$\mathrm{pass}^{\mathrm{true}}\,@k = 1-(1-p)^k. \tag{90}$$

Proposition 10 (False-positive amplification)**.** *If $f > 0$, then even a question with no true correct solution probability, $p = 0$, has*

$$\mathrm{pass}^{\mathrm{obs}}@k = 1 - (1 - f)^k \to 1, \qquad \mathrm{pass}^{\mathrm{true}}@k = 0. \tag{91}$$

More generally, for small acceptance perturbations $\delta = a - p$, the first-order distortion is

$$\mathrm{pass}^{\mathrm{obs}}@k - \mathrm{pass}^{\mathrm{true}}@k = k(1 - p)^{k-1}\delta + O(\delta^2). \tag{92}$$

Thus tiny per-sample verifier errors are magnified at budgets of order $1/p$ or $1/f$.

Proof. Equation (91) follows by substituting $p = 0$ into (88) and (89). The expansion is the Taylor expansion of $g(x) = 1 - (1 - x)^k$ around $x = p$. Q.E.D.

The second problem appears when a downstream system must return one candidate. Consider the natural policy that returns a uniformly random accepted candidate if at least one candidate is accepted, and otherwise returns nothing. Conditional on the number of accepted candidates being positive, the accepted candidates have precision

$$\pi_{\mathrm{acc}} = \Pr(Y = 1 \mid V = 1) = \frac{rp}{a}, \qquad a > 0. \tag{93}$$

Proposition 11 (Selector gap under an imperfect verifier)**.** *Under the binary verifier model above, the probability that the returned accepted candidate is truly correct is*

$$S_k = \left\{1 - (1 - a)^k\right\} \frac{rp}{a}, \tag{94}$$

while the probability that the returned candidate is a false positive is

$$F_k = \left\{1 - (1 - a)^k\right\} \frac{f(1 - p)}{a}. \tag{95}$$

Consequently,

$$\lim_{k\to\infty} S_k = \frac{rp}{rp + f(1 - p)}, \qquad \lim_{k\to\infty} F_k = \frac{f(1 - p)}{rp + f(1 - p)}. \tag{96}$$

The reported accepted $\mathrm{pass}@k$ converges to one whenever $a > 0$, but returned correctness converges only to verifier precision.

Proof. Each candidate falls into one of three relevant categories: true accepted with probability rp, false accepted with probability $f(1 - p)$, or not accepted with probability $1 - a$. Let L be the number of accepted candidates. Conditional on $L > 0$, the accepted candidates are an i.i.d. multinomial subsample with true-accepted fraction rp/a in expectation; uniform selection among accepted candidates therefore returns a true candidate with probability rp/a. Multiplying by $\Pr(L > 0) = 1 - (1 - a)^k$ gives (94). The false-positive expression is identical with $f(1 - p)$ in place of rp. Taking $k \to \infty$ proves (96). Q.E.D.

Table 50 shows the scale of the effect. With $p = 0.01$ and a false-positive rate $f = 0.01$, reported accepted $\mathrm{pass}@100$ is 0.866, but the probability that the returned accepted sample is truly correct is only 0.435; the other 0.431 mass is returned false positives. If $p = 0$, even a tiny false-positive rate $f = 0.001$ creates reported $\mathrm{pass}@100 = 0.095$ despite zero true correctness.

Random model-pair check. The random model-pair check samples 200,000 pairs of synthetic models with

$$p \sim \mathrm{Unif}(0, 0.1), \qquad r \sim \mathrm{Unif}(0.9, 1), \qquad f \sim \mathrm{LogUnif}(10^{-5}, 0.05),$$

independently for each model. For each pair, we compared the ranking induced by reported accepted $\mathrm{pass}@k$ in (89) with the ranking induced by returned true correctness S_k in (94). The Monte Carlo standard error of each displayed disagreement probability is at most $\sqrt{0.25/200000} \approx 0.0011$.

The disagreement probability grows with k, reaching 0.409 at $k = 250$ in this synthetic family. The reason is that reported $\mathrm{pass}@k$ rewards batch-level acceptance, while deployment correctness also depends on the precision of the accepted set. As k grows, acceptance becomes almost guaranteed, so the remaining distinction is accepted-candidate precision.

Table 50: Verifier and selection gap under a binary imperfect verifier. Here p is the true per-sample correctness probability, r is the true-positive rate, f is the false-positive rate, and selection is uniform among accepted samples. All rows set $r = 1$. Reported $\mathrm{pass@}k$ is the probability that at least one sample is accepted by the verifier; selected-true is the probability that the returned accepted sample is truly correct; selected-false is the probability that the returned sample is a false positive.

Scenario	p	f	k	Reported	Selected true	Selected false
Impossible task, tiny FPR	0.000	0.001	1	0.001	0.000	0.001
	0.000	0.001	10	0.010	0.000	0.010
	0.000	0.001	100	0.095	0.000	0.095
Rare true success, low FPR	0.010	0.001	1	0.011	0.010	0.001
	0.010	0.001	10	0.105	0.095	0.009
	0.010	0.001	100	0.669	0.609	0.060
Rare true success, matched FPR	0.010	0.010	1	0.020	0.010	0.010
	0.010	0.010	10	0.182	0.092	0.091
	0.010	0.010	100	0.866	0.435	0.431
Moderate true success, FPR	0.050	0.010	1	0.060	0.050	0.010
	0.050	0.010	10	0.459	0.385	0.073
	0.050	0.010	100	0.998	0.839	0.159

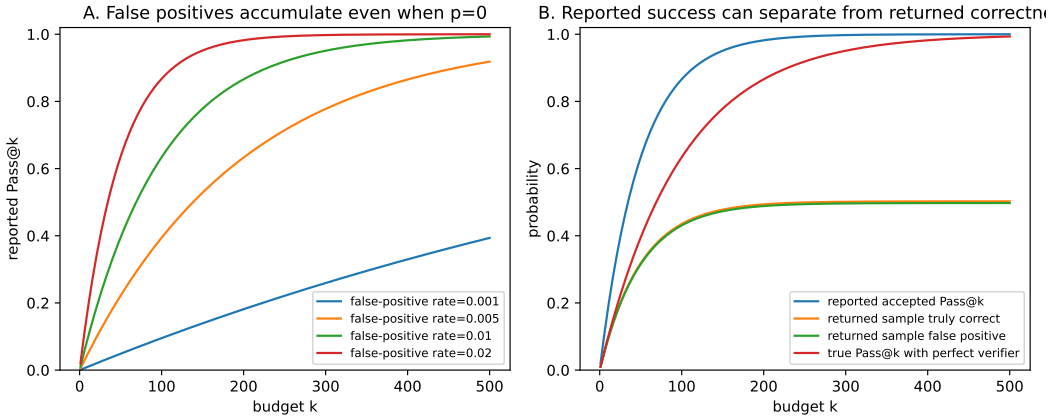

Figure 36: Verifier and selector fragility. Panel A shows reported accepted $\mathrm{pass@}k$ for impossible tasks ($p = 0$) under different per-sample false-positive rates. Panel B fixes $p = 0.01$, $r = 1$, and $f = 0.01$: reported accepted $\mathrm{pass@}k$ approaches one, but the probability of returning a truly correct accepted sample saturates near verifier precision.

Table 51: Random model-pair check for verifier/selector mismatch. We sampled $p \sim \mathrm{Unif}(0, 0.1)$, $r \sim \mathrm{Unif}(0.9, 1)$, and $f \sim \mathrm{LogUnif}(10^{-5}, 0.05)$ independently for each model. The table reports the fraction of 200,000 pairs for which the ranking by reported accepted $\mathrm{pass@}k$ disagrees with the ranking by selected true correctness.

k	1	5	10	25	50	100	250
Disagreement	0.079	0.089	0.101	0.142	0.207	0.302	0.409

Practical implication. $\mathrm{pass@}k$ is an oracle-discovery metric. It is appropriate when the verifier is effectively exact and oracle selection is available. When the verifier is imperfect, one should report false-positive audits, verifier precision/recall, or selected-candidate correctness in addition to accepted $\mathrm{pass@}k$. Otherwise, repeated sampling can convert small per-sample verifier errors into large apparent gains.

L.2 Dependence & Effective Sample-Size Collapse

The i.i.d. interpretation of $\mathrm{pass@}k$ treats the budget axis as independent repeated sampling. The finite-bank estimator in (46) remains a valid within-bank discovery statistic, but its interpretation as test-time scaling depends on how the k attempts are generated. Correlated attempts, duplicated reasoning traces, shared latent decoding states, or prompt-level randomness can make a nominal budget of k behave like far fewer independent attempts. The resulting error is budget-axis miscalibration, distinct from the large-budget convergence failures above.

A non-identification fact. For one question, let $X_j \in \{0,1\}$ denote correctness of attempt j and assume only that each attempt has the same marginal success probability $\Pr(X_j = 1) = p$. Define

$$D_k = \Pr\left(\max_{1\le j\le k} X_j = 1\right). \tag{97}$$

Then the marginal p leaves D_k unidentified. The sharp Fréchet bounds are

$$p \le D_k \le \min\{1, kp\}. \tag{98}$$

The lower bound is achieved when all attempts are perfectly coupled, $X_1 = \cdots = X_k$; the upper bound is attained by disjoint success events when $kp \le 1$ and by events covering the probability space when $kp \ge 1$. The independent formula $1-(1-p)^k$ is only one point inside this interval.

A common positive-dependence model is exchangeability with a latent per-batch success probability. Let

$$X_j \mid P \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(P), \qquad j = 1,\dots,k, \qquad \mathbb{E}[P] = p, \tag{99}$$

where P is a latent random variable. This is the de Finetti representation for an infinitely exchangeable Bernoulli sequence (Gelman et al., 2013, Sec. 5.2).

Proposition 12 (Positive dependence depresses discovery)**.** *For $k \ge 2$ under* (99)*,*

$$D_k = 1 - \mathbb{E}[(1-P)^k] \le 1 - (1-p)^k, \tag{100}$$

with equality if and only if $P = p$ almost surely. Thus any non-degenerate latent variation makes the repeated-sampling gain smaller than the i.i.d. $\mathrm{pass@}k$ *curve with the same* $\mathrm{pass@}1$*.*

Proof. Conditional on P, the probability that all k attempts fail is $(1-P)^k$, so $D_k = 1-\mathbb{E}[(1-P)^k]$. For $k \ge 2$, the map $x \mapsto (1-x)^k$ is strictly convex on $[0,1]$. Jensen's inequality gives $\mathbb{E}[(1-P)^k] \ge (1-\mathbb{E}[P])^k = (1-p)^k$, with equality only when P is degenerate. Q.E.D.

Effective independent budget. For $0 < p < 1$ and $p \le D_k < 1$, define the independent-equivalent budget

$$k_{\text{eff}}(D_k; p) = \frac{\log(1-D_k)}{\log(1-p)}. \tag{101}$$

This is the number of independent attempts that would give the same discovery probability. Under independence, $k_{\text{eff}} = k$; under positive dependence, Proposition 12 implies $k_{\text{eff}} \le k$.

The numerical check uses a beta-binomial exchangeable family. Let

$$P \sim \mathrm{Beta}(p\tau, (1-p)\tau), \qquad \rho = \frac{1}{\tau+1}, \tag{102}$$

where ρ is the intraclass correlation among attempts. Then

$$D_k(\rho) = 1 - \frac{\mathrm{B}(p\tau, (1-p)\tau + k)}{\mathrm{B}(p\tau, (1-p)\tau)}. \tag{103}$$

The endpoint $\rho = 0$ is the i.i.d. curve and $\rho \to 1$ approaches perfect lockstep dependence, $D_k = p$.

Table 52 fixes $p = 0.05$ and nominal $k = 64$. A modest intraclass correlation $\rho = 0.10$ reduces the discovery probability from 0.962 to 0.623, equivalent to only 19.0 independent attempts. At $\rho = 0.25$, the same nominal budget behaves like fewer than ten independent attempts.

Table 52: Effective sample-size collapse under exchangeable dependence. The marginal success probability is fixed at $p = 0.05$ and the nominal budget is $k = 64$. The intraclass correlation ρ is implemented by a beta-binomial de Finetti mixture; k_{eff} is the independent-sampling budget that would give the same discovery probability.

ρ	Discovery probability	k_{eff}	$k_{\text{eff}}/64$
0.00	0.962	64.0	1.000
0.01	0.920	49.3	0.770
0.05	0.763	28.1	0.438
0.10	0.623	19.0	0.297
0.25	0.391	9.7	0.151
0.50	0.213	4.7	0.073
1.00	0.050	1.0	0.016

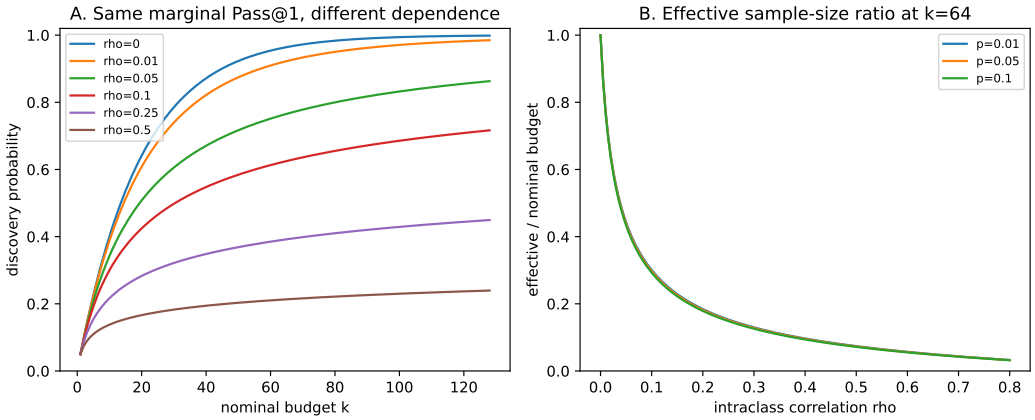

Figure 37: Dependence-induced effective sample-size collapse. Panel A fixes marginal $\text{pass@1} = p = 0.05$ and plots the discovery curve under beta-binomial exchangeable dependence. Panel B plots the independent-equivalent budget ratio k_{eff}/k at nominal $k = 64$. Even moderate positive dependence can erase much of the apparent repeated-sampling budget.

Practical implication. A pass@k report should specify the sampling protocol and avoid extrapolating from pass@1 or a small set of budgets using an i.i.d. formula unless independence has been checked. Diagnostics include duplicate rates, pairwise agreement of correctness indicators, trace diversity, and an empirical effective-budget curve. Without such checks, a nominal k may substantially overstate the amount of independent search performed by the model.

L.3 One-Sided Utility & Harmful-Attempt Blindness

pass@k is a one-sided discovery utility, therefore it gives credit for success but assigns no negative value to incorrect responses. This is appropriate for a narrow offline discovery experiment in which failed samples are discarded at zero cost. It is incomplete when attempts consume resources, require human review, trigger tool calls, or can produce harmful outputs. Agent benchmarks and interactive tool-use settings make this issue especially salient, because failed attempts may change the environment even when final reporting keeps only the response bank (Yao et al., 2025). The batch utility is one when at least one response is correct and zero otherwise, with no separate penalty for incorrect responses.

Setup. Let each sampled response have one of three labels:

$$Z_j \in \{C, H, B\},$$

where C means correct, H means harmful or costly, and B means benign incorrect. Write

$$p_C = \Pr(Z = C), \qquad p_H = \Pr(Z = H), \qquad p_C + p_H \le 1. \tag{104}$$

For i.i.d. attempts,

$$\text{pass@}k = \Pr(\exists j : Z_j = C) = 1 - (1 - p_C)^k, \tag{105}$$

while the probability of at least one harmful attempt is

$$H@k = \Pr(\exists j : Z_j = H) = 1 - (1 - p_H)^k. \tag{106}$$

Proposition 13 ($\text{pass@}k$ is blind to harmful-attempt probability)**.** *For every fixed p_C and budget k, all values of $p_H \in [0, 1 - p_C]$ give exactly the same $\text{pass@}k$ value in* (105)*. However, the harmful-attempt probability ranges from* 0 *to*

$$1 - p_C^k \tag{107}$$

when p_H ranges from 0 *to* $1 - p_C$*.*

Proof. Equation (105) depends only on p_C, so changing p_H while holding p_C fixed leaves $\text{pass@}k$ unchanged. Equation (106) is increasing in p_H, and at the largest feasible value $p_H = 1 - p_C$ it equals $1 - p_C^k$. Q.E.D.

Thus $\text{pass@}k$ assigns the same value to a model whose incorrect attempts are harmless and a model whose incorrect attempts are all harmful. A simple utility-aware metric with harm cost $\lambda > 0$ would be

$$U_\lambda(k) = \Pr(\exists j : Z_j = C) - \lambda \Pr(\exists j : Z_j = H) = 1 - (1 - p_C)^k - \lambda\{1 - (1 - p_H)^k\}. \tag{108}$$

$\text{pass@}k$ is the special case $\lambda = 0$.

Table 53: Harmful-attempt blindness at fixed per-sample correctness $p_C = 0.05$. $\text{pass@}k$ is identical across rows with the same k, but the probability of at least one harmful attempt changes rapidly with the per-sample harmful-output probability p_H. The final column shows a simple utility $U_k = \Pr(\exists\, C) - \Pr(\exists\, H)$.

p_H	k	$\text{pass@}k$	$\Pr(\exists\,\text{harm})$	U_k
0.00	1	0.050	0.000	0.050
	10	0.401	0.000	0.401
	50	0.923	0.000	0.923
	100	0.994	0.000	0.994
0.01	1	0.050	0.010	0.040
	10	0.401	0.096	0.306
	50	0.923	0.395	0.528
	100	0.994	0.634	0.360
0.05	1	0.050	0.050	0.000
	10	0.401	0.401	0.000
	50	0.923	0.923	0.000
	100	0.994	0.994	0.000
0.10	1	0.050	0.100	-0.050
	10	0.401	0.651	-0.250
	50	0.923	0.995	-0.072
	100	0.994	1.000	-0.006

Table 53 fixes $p_C = 0.05$. $\text{pass@}100$ is 0.994 regardless of harmful-output probability. But when $p_H = 0.01$, the chance of at least one harmful attempt is already 0.634; when $p_H = 0.10$, it is essentially one. Under the utility $U_k = \Pr(\exists C) - \Pr(\exists H)$, large-budget sampling can be unattractive even though $\text{pass@}k$ is near one.

Practical implication. When failed attempts are truly discarded at zero cost, this failure mode may be irrelevant. When attempts have side effects, human-review cost, monetary cost, latency cost, or safety risk, $\text{pass@}k$ should be paired with negative-outcome rates and an explicit utility or risk model. Otherwise, repeated sampling can improve the reported discovery score while increasing the probability of at least one bad event.

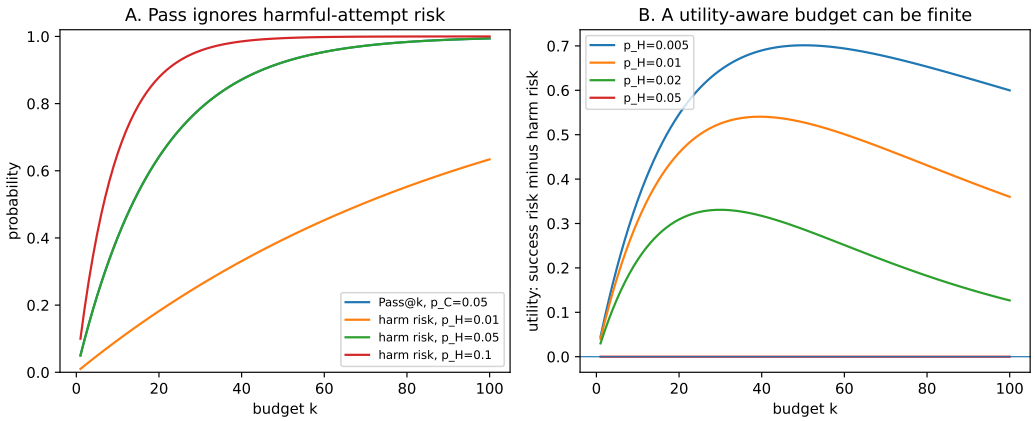

Figure 38: One-sided utility failure. Panel A fixes $p_C = 0.05$: pass@k is unchanged as p_H varies, but harmful-attempt risk grows with both p_H and k. Panel B plots the utility $U_k = \Pr(\exists C) - \Pr(\exists H)$; for nonzero harm rates, the utility-aware budget can differ sharply from the budget favored by pass@k.

L.4 Finite-bank Uncertainty & False Ranking

Even with the target, verifier, utility, and sampling protocol fixed, the finite-bank pass@k estimator is a statistic with nontrivial sampling uncertainty. The combinatorial formula averages over all $\binom{N}{k}$ subsets of a bank, but those subsets are highly dependent because they reuse the same N generated candidates. The subset count can greatly exceed the amount of independent evidence. Bayesian analyses of pass@k emphasize that limited numbers of trials can yield unstable rankings and that uncertainty intervals should be part of the evaluation protocol (Hariri et al., 2026c).

Setup. For one homogeneous question, let

$$X_1, \dots, X_N \overset{\text{iid}}{\sim} \text{Bernoulli}(p), \qquad C = \sum_{j=1}^{N} X_j.$$

The finite-bank row estimator is the complete U-statistic

$$\widehat{Q}_{N,k} = 1 - \frac{\binom{N-C}{k}}{\binom{N}{k}} = \frac{1}{\binom{N}{k}} \sum_{A \subseteq [N]:|A|=k} \mathbf{1}\left\{\max_{j \in A} X_j = 1\right\}. \tag{109}$$

It is unbiased for

$$Q_k = 1 - (1-p)^k. \tag{110}$$

Proposition 14 (Exact row-level variance)**.** *Let A and B be two independent uniformly random k-subsets of $\{1, \dots, N\}$, and let $S = |A \cap B|$. Then*

$$\text{Var}(\widehat{Q}_{N,k} \mid p) = \mathbb{E}\left[(1-p)^{2k-S}\right] - (1-p)^{2k}, \tag{111}$$

where

$$\Pr(S = s) = \frac{\binom{k}{s}\binom{N-k}{k-s}}{\binom{N}{k}}, \qquad s = \max\{0, 2k-N\}, \dots, k. \tag{112}$$

Equivalently,

$$\text{Var}(\widehat{Q}_{N,k} \mid p) = \sum_{s=\max\{0,2k-N\}}^{k} \frac{\binom{k}{s}\binom{N-k}{k-s}}{\binom{N}{k}} (1-p)^{2k-s} - (1-p)^{2k}. \tag{113}$$

Proof. Let

$$F_A = \mathbf{1}\{X_j = 0 \text{ for every } j \in A\}.$$

Then $\widehat{Q}_{N,k} = 1 - \binom{N}{k}^{-1} \sum_A F_A$. For two subsets A, B, both failure indicators equal one exactly when every element in $A \cup B$ is a failure. Since $|A \cup B| = 2k - |A \cap B|$, this probability is $(1-p)^{2k-S}$. Averaging over the random overlap S gives $\mathbb{E}[\bar{F}^2]$ for $\bar{F} = \binom{N}{k}^{-1} \sum_A F_A$. Since $\mathrm{Var}(\widehat{Q}_{N,k}) = \mathrm{Var}(\bar{F})$ and $\mathbb{E}[\bar{F}] = (1-p)^k$, subtracting $(1-p)^{2k}$ yields (111). The overlap distribution is hypergeometric, which gives (112) and (113). Q.E.D.

Two endpoint checks are useful. At $k = 1$,

$$\mathrm{Var}(\widehat{Q}_{N,1} \mid p) = \frac{p(1-p)}{N}, \tag{114}$$

which is the usual sample-mean variance. At $k = N$,

$$\widehat{Q}_{N,N} = \mathbf{1}\{C > 0\}, \qquad \mathrm{Var}(\widehat{Q}_{N,N} \mid p) = Q_N(1 - Q_N), \tag{115}$$

so this statistic retains only the coverage indicator $\mathbf{1}\{C > 0\}$ from the N observations.

Table 54: Exact finite-bank uncertainty for homogeneous rows with $N = 80$ generated samples per question, $M = 100$ questions, and marginal success probability $p = 0.05$. The row standard deviation is for the unbiased finite-bank pass@k row estimator; the dataset standard error is the row standard deviation divided by $\sqrt{M}$.

k	Target pass@k	Row SD	Dataset SE
1	0.050	0.024	0.002
5	0.226	0.099	0.010
10	0.401	0.156	0.016
20	0.642	0.195	0.020
40	0.871	0.173	0.017
80	0.983	0.127	0.013

Table 54 reports exact standard errors for $N = 80$, $M = 100$, and $p = 0.05$. The row-level standard deviation is appreciable at intermediate budgets. For fixed heterogeneous questions with independent response banks, the dataset variance is $M^{-2} \sum_i \mathrm{Var}(\widehat{Q}_{N,k} \mid p_i)$. An additional between-question term arises only when the questions themselves are sampled from a population.

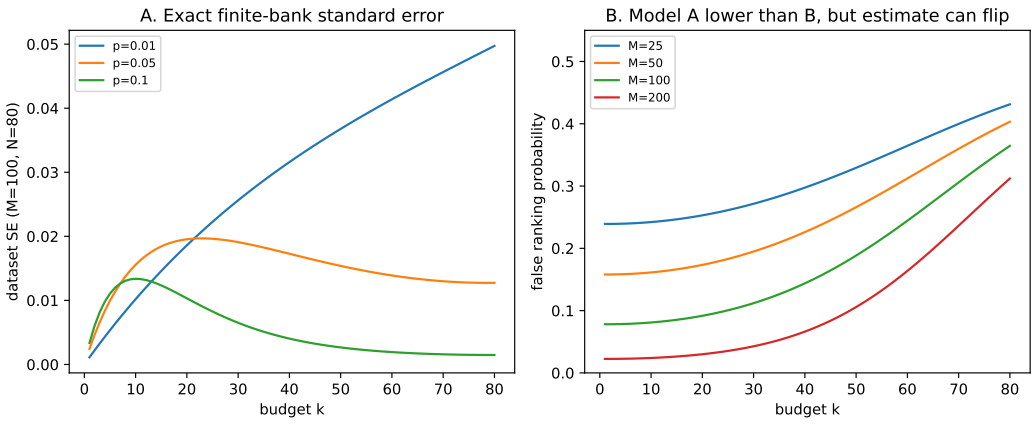

Figure 39: Finite-bank uncertainty. Panel A gives exact dataset standard errors from Proposition 14 for $N = 80$ and $M = 100$. Panel B considers two homogeneous models with $p_A = 0.050$ and $p_B = 0.055$; model B has the larger population pass@k for every k, but finite-bank estimates can still rank A above B. The plotted probabilities use a normal approximation. At $k = N = 80$, the exact strict-reversal probability is 0.237 even for $M = 200$ (Table 55); the approximation overstates reversals near saturation.

False-ranking calculation. For two homogeneous models A and B with independent response banks across models and questions, a normal approximation gives

$$\Pr\{\text{pass}^A @k > \text{pass}^B @k\} \approx \Phi\left(-\frac{Q_k^B - Q_k^A}{\sqrt{\{\text{Var}_A(\widehat{Q}_{N,k}) + \text{Var}_B(\widehat{Q}_{N,k})\}/M}}\right), \tag{116}$$

where Φ is the standard normal distribution function. Table 55 uses the exact row variances from Proposition 14 with $N = 80$, $p_A = 0.050$, and $p_B = 0.055$.

Table 55: Strict false-ranking probabilities for independent homogeneous model banks with $p_A = 0.050$, $p_B = 0.055$, and $N = 80$. The first four probability columns use normal approximations with exact variances. The last column uses the exact binomial coverage calculation at $k = 80$, excluding ties.

M	$k = 1$	$k = 10$	$k = 40$	$k = 80$ (normal)	$k = 80$ (exact)
25	0.239	0.242	0.297	0.431	0.166
50	0.158	0.161	0.226	0.403	0.223
100	0.078	0.081	0.144	0.365	0.248
200	0.022	0.024	0.066	0.312	0.237

At $k = N$, the number of covered questions for model a is $B_a \sim \text{Binomial}(M, Q_N^a)$. Under independent model banks, the exact strict-reversal probability is

$$\Pr(B_A > B_B) = \sum_{x=0}^{M} \Pr(B_A = x) \Pr(B_B \le x - 1).$$

The discrete calculation matters near saturation, where ties have substantial probability. At $M = 25$, the normal approximation gives 0.431 but the exact probability is 0.166; at $M = 200$, the corresponding values are 0.312 and 0.237. The point-estimate ranking remains uncertain, but the normal approximation overstates strict reversals in these examples.

Finite-bank ranking uncertainty is one motivation for $\text{TailPass}@k$, which reports uncertainty in both the discovery–stability profile and its scalar summaries (Section 3.1). The pairwise comparison rule in Appendix J.4 leaves an ordering unresolved when the 95% credible interval for the score difference contains zero. This failure mechanism is distinct from the deterministic budget-dependent reversals in Section K.1: sampling noise can reverse an empirical ordering even when the population ranking is stable for all k.

M Failure Modes of Threshold & Area Metrics for Repeated Sampling

Scalar repeated-sampling metrics select coordinates or averages of the threshold tail surface. Let

$$t_\tau(k) = \begin{cases} 1, & \tau = 0, \\ \lceil \tau k \rceil, & 0 < \tau \le 1, \end{cases}$$

and recall the population coordinate $S_{k,t} = M^{-1} \sum_i \Pr\{\text{Binomial}(k, p_i) \ge t\}$ and the normalized profile $S_k(\rho) = S_{k,\lceil k\rho \rceil}$ for $\rho \in (0, 1]$, with $S_k(0^+) = S_{k,1}$. The finite-bank analogues replace $S_{k,t}$ by $\widehat{S}^{\text{fs}}_{k,t}$ in Equation (44), for reporting budgets no larger than the observed bank. In this notation, thresholded G-Pass@k_τ is $S_{k,t_\tau(k)}$, with the discovery endpoint represented by $S_k(0^+) = S_{k,1}$; the strict endpoint G-Pass@k_1 coincides with pass^k; strict binary Maj@k is $S_{k,\lfloor k/2 \rfloor + 1}$; mG-Pass@$k$ is the upper-half threshold utility

$$2 \int_{1/2}^{1} S_k(\rho)\, d\rho = \sum_{t=1}^{k} \omega_{t,k}^{\text{upper}} S_{k,t}, \qquad \omega_{t,k}^{\text{upper}} = 2\ \text{Leb}(C_{t,k} \cap (1/2, 1]),$$

where Leb is Lebesgue measure and $C_{1,k} = [0, 1/k]$, $C_{t,k} = ((t-1)/k, t/k]$ for $t \ge 2$; and AUC@K is the uniform discrete budget average $K^{-1} \sum_{\ell=1}^{K} S_{\ell,1}$. This convention also applies to

the response-bank diagnostics; the convergence and rank-correlation experiment uses the trapezoidal weights in Equation (41). G-Pass@k_τ and mG-Pass@k were introduced to evaluate both potential and stability of repeated reasoning (Liu et al., 2025), while pass^k measures all-success reliability in agent evaluation (Yao et al., 2025). Majority-based selection is closely related to self-consistency decoding, which aggregates multiple sampled reasoning paths by a voting rule (Wang et al., 2023). AUC@K is a uniform utility over discovery budgets.

Table 56: Extended repeated-sampling metrics as scalar views of the discovery–stability profile. The listed failures are specific to the scalar view; all metrics also inherit verifier, dependence, utility, and finite-bank uncertainty failures from Appendix L.

Metric	Profile functional	Metric-specific failure diagnostic
G-Pass@k_τ	$S_{k,t_\tau(k)}$, with $t_0(k)=1$ and $t_\tau(k)=\lceil \tau k \rceil$ for $0<\tau\le 1$.	A hard threshold aliases all below-threshold margins and all above-threshold margins; rankings can reverse as τ changes even at fixed k.
pass^k / G-Pass@k_1	$S_{k,k}$, the all-correct endpoint.	Increasing k simultaneously adds attempts and tightens the reliability requirement; under dependence, strict-stability scores can be inflated by batch-level correlation.
Maj@k	$S_{k,\lfloor k/2\rfloor+1}$ for strict binary majority, or a modal-answer selection rule when answers are clustered.	The binary version is a single threshold coordinate with parity and tie-convention effects; the modal-answer version is not identifiable from binary correctness counts.
mG-Pass@k	$2\int_{1/2}^{1} S_k(\rho)\,d\rho = \sum_t \omega_{t,k}^{\mathrm{upper}} S_{k,t}$.	The high-threshold area gives asymptotic zero credit to questions with $p_i \le 1/2$ and can hide very different majority-to-unanimity decay shapes.
AUC@K	$K^{-1}\sum_{\ell=1}^{K} S_{\ell,1}$ for a uniform budget grid.	The score is a budget-weighted discovery utility; it can disagree with the operating budget and changes when the upper limit or budget grid changes.

The scalar-projection failures in Appendix K apply beyond the threshold $t=1$. Any fixed coordinate $S_{k,t}$, any threshold area, and any budget area can reverse rankings, leave large equivalence classes, and target a functional different from average accuracy. The auxiliary assumption failures in Appendix L also remain active. Verifier errors change the observed tail surface from true correctness to accepted correctness. Dependence changes the meaning of the budget axis. Harmful or costly attempts are invisible to binary tails. Finite banks require uncertainty intervals before close rankings are treated as decisive. Metric-specific failures arise from the particular threshold, majority, or area structure of the extended metrics.

M.1 Large-budget Targets Depend on the Chosen Threshold

For normalized thresholds, a large budget estimates a survival curve of per-question success probabilities. The endpoint chosen by the metric determines which survival functional is approached.

Proposition 15 (Large-budget targets of extended metrics)**.** *Let $F_M = M^{-1}\sum_{i=1}^{M}\delta_{p_i}$ be the empirical distribution of question-level success probabilities. Assume independent attempts conditional on each p_i. For any $\tau\in(0,1)$ with $F_M\{\tau\}=0$,*

$$\lim_{k\to\infty} S_{k,t_\tau(k)} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}\{p_i>\tau\}. \tag{117}$$

The strict endpoint satisfies

$$\lim_{k\to\infty} S_{k,k} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}\{p_i=1\}, \tag{118}$$

and strict binary majority satisfies

$$\lim_{k\to\infty} S_{k,\lfloor k/2\rfloor+1} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{p_i > 1/2\} \tag{119}$$

when $F_M\{1/2\} = 0$*. The high-threshold area obeys*

$$\lim_{k\to\infty} 2\int_{1/2}^{1} S_k(\rho)\, d\rho = \frac{2}{M}\sum_{i=1}^{M}(p_i - 1/2)_+, \tag{120}$$

while the normalized discovery area satisfies

$$\lim_{K\to\infty} \frac{1}{K}\sum_{\ell=1}^{K} S_{\ell,1} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{p_i > 0\}. \tag{121}$$

Proof. For $0 < \tau < 1$, the weak law of large numbers gives $\Pr\{\mathrm{Binomial}(k,p)/k \geq t_\tau(k)/k\} \to \mathbf{1}\{p > \tau\}$ at $p \neq \tau$. Averaging over the finite set of questions gives (117). The strict endpoint is $q_{k,k}(p) = p^k$, which converges to $\mathbf{1}\{p = 1\}$. Majority is the special case with normalized threshold tending to $1/2$. Equation (120) follows by applying dominated convergence under the threshold integral and then using $\int_{1/2}^{1} \mathbf{1}\{p_i > \rho\}\, d\rho = (p_i - 1/2)_+$. For the discovery budget area, the pointwise Cesàro mean of $1-(1-p_i)^\ell$ over $\ell = 1,\ldots,K$ converges to $\mathbf{1}\{p_i > 0\}$; averaging over questions gives (121) (Billingsley, 1995). Q.E.D.

These limits are well specified once the target is stated explicitly. They become failure modes when a high-threshold or area metric is interpreted as a generic measure of model quality. G-Pass@k_τ approaches the mass of questions with $p_i > \tau$; pass^k approaches the mass of perfectly reliable questions with $p_i = 1$; mG-Pass@k ignores all questions whose limiting success probability is at most one half; and AUC@K approaches coverage, just as large-budget pass@k does. These targets generally differ from average accuracy, although equality can occur for particular success-probability distributions.

M.2 Threshold-induced Ranking Reversals

A thresholded metric changes the evaluation target even when the sampling budget is fixed. The derivative identity in Equation (9) shows why: changing t moves the sensitivity of $q_{k,t}(p)$ to a different region of the success-probability axis. Low thresholds reward broad rare success. High thresholds reward a smaller set of questions that are solved repeatedly.

Table 57: Threshold-induced reversal at fixed budget $k = 16$. The broad model has $p_i = 0.40$ on every question; the narrow model has $p_i = 0.90$ on 35% of questions and $p_i = 0$ otherwise. The broad model has larger average accuracy (0.400 versus 0.315), but high-threshold G-Pass@$16_{0.5}$ and mG-Pass@16 favor the narrow reliable subset.

τ	$t = \lceil \tau k \rceil$	Broad moderate	Narrow reliable	Larger score
0.00	1	1.000	0.350	Broad moderate
0.25	4	0.935	0.350	Broad moderate
0.50	8	0.284	0.350	Narrow reliable
0.75	12	0.005	0.344	Narrow reliable
1.00	16	0.000	0.065	Narrow reliable
mG-Pass@16	–	0.028	0.280	Narrow reliable
Average accuracy	–	0.400	0.315	Broad moderate

Table 57 gives a population calculation at $k = 16$. The broad model has higher average accuracy and dominates at low thresholds. The narrow model solves only 35% of questions with nonzero probability, but it solves those questions reliably; it therefore wins at the majority, high-threshold, and all-correct coordinates. The reversal follows from choosing a different coordinate of the same tail surface.

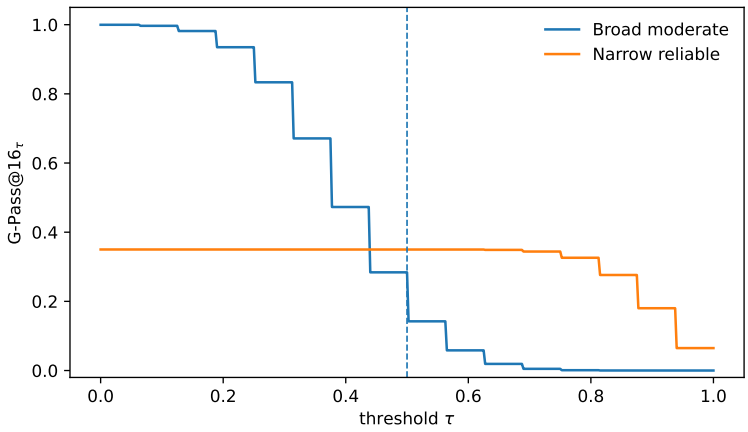

Figure 40: Threshold-induced ranking reversal at fixed budget $k = 16$. The broad model wins under discovery-oriented G-Pass@k_τ thresholds; the narrow reliable model wins once the threshold requires majority-level or stronger repeatability. A single G-Pass@k_τ value encodes an implicit reliability requirement.

The same issue appears at the strict endpoint. Increasing k in pass^k requires every additional attempt to be correct. A model with per-question success probability $p < 1$ has $p^k \to 0$ at that endpoint. Thus a decreasing pass^k curve need not indicate that the model becomes worse when more samples are drawn. It may only reflect that the reported event has become more stringent.

M.3 Dependence Can Inflate Stability Coordinates

For pass@k, exchangeable positive dependence depresses discovery relative to an i.i.d. model with the same marginal success probability; Proposition 12 uses Jensen's inequality for the convex failure probability. High-threshold coordinates have no single dependence direction. Let $Y_j \mid P \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(P)$ with $\mathbb{E}[P] = p$ (Gelman et al., 2013, Sec. 5.2), and let $X_k = \sum_{j=1}^{k} Y_j$. Then

$$\Pr\{X_k \geq t\} = \mathbb{E}\big[q_{k,t}(P)\big]. \tag{122}$$

For $t = k$, the map $p \mapsto p^k$ is convex, so $\mathbb{E}[P^k] \geq (\mathbb{E}P)^k$. Batch-level dependence can therefore inflate strict stability even while it reduces low-threshold discovery. For intermediate thresholds, the sign depends on the marginal p and the chosen coordinate: dependence moves probability mass toward all-success and all-failure batches, so it can raise a tail coordinate when the independent binomial tail is below the perfectly correlated limit p, and lower it when the independent tail is above that limit.

Table 58: Dependence can change high-threshold metrics in either direction. The beta-binomial de Finetti mixture fixes the marginal success probability p and varies intraclass correlation ρ. For majority at $k = 17$, $t = 9$; for strict stability, $k = t = 8$ and $p = 0.20$.

ρ	Maj@17, $p = 0.45$	Maj@17, $p = 0.55$	pass^8, $p = 0.20$
0.00	0.337	0.663	0.000
0.05	0.378	0.622	0.000
0.10	0.398	0.602	0.000
0.25	0.426	0.574	0.006
0.50	0.443	0.557	0.041
1.00	0.450	0.550	0.200

This direction change is specific to threshold metrics. A nominal G-Pass@k_τ, Maj@k, or pass^k score can conflate genuine per-attempt reliability with batch correlation induced by shared prompts, common decoding states, duplicated reasoning traces, or near-deterministic answer templates. Dependence diagnostics therefore need to be threshold-specific: an effective-budget correction that is appropriate for pass@k can fail for stability coordinates.

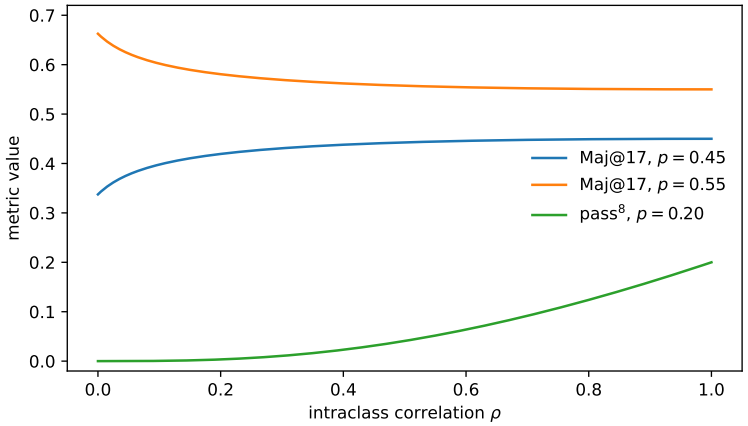

Figure 41: Direction of dependence bias for high-threshold metrics under a beta-binomial exchangeable family. Majority at $p = 0.45$ increases with intraclass correlation because correlated batches sometimes move many attempts above the majority threshold together. Majority at $p = 0.55$ decreases because independent sampling already benefits from concentration around a mean above the threshold. The strict endpoint pass^8 is convex in p and is inflated by positive dependence.

M.4 Majority Voting Requires Answer-Cluster Information

Strict binary $\mathrm{Maj@}k$ is a threshold coordinate of the correctness-count tail. Many practical majority or self-consistency evaluations, however, select a modal answer string or a modal normalized answer and then check whether the selected answer is correct (Wang et al., 2023). That selection rule depends on the partition of responses into answer-equivalence classes as well as the binary correctness vector. The binary response matrix R is therefore insufficient for auditing modal-answer $\mathrm{Maj@}k$.

Table 59: Majority-vote selection is not identified by binary correctness counts. The symbols C_a denote distinct verified-correct surface forms and W_a distinct incorrect surface forms. Both first two rows have the same binary correctness vector $(1, 1, 1, 0, 0)$ and the same binary $\mathrm{Maj@5}$ value, but exact-string modal voting returns different correctness unless correct answers are canonicalized before voting.

Response multiset	Binary correctness vector	Binary $\mathrm{Maj@5}$	Modal-answer selection
$\{C_1, C_2, C_3, W, W\}$	$(1, 1, 1, 0, 0)$	1	incorrect
$\{C, C, C, W_1, W_2\}$	$(1, 1, 1, 0, 0)$	1	correct
$\{C, C, W_1, W_2, W_3\}$	$(1, 1, 0, 0, 0)$	0	correct under plurality

Table 59 separates the two notions. The first two rows have identical binary correctness counts and identical binary $\mathrm{Maj@5}$ values. Exact-string modal voting nevertheless selects a wrong answer in the first row and a correct answer in the second. The third row shows the opposite phenomenon: a correct answer can win a plurality even when fewer than a strict majority of samples are correct, provided the wrong answers are fragmented. Majority-vote evaluation must therefore report the canonicalization rule, the tie-breaking rule, and whether the vote is over exact strings, semantic equivalence classes, final answers, programs, or verifier outcomes. Reproducing the metric requires the answer clusters as well as the binary correctness labels.

M.5 Area Metrics Hide Where the Area Comes From

$\mathrm{mG\text{-}Pass@}k$ and $\mathrm{AUC@}K$ reduce a curve to an area. Such reductions are useful when the integration measure is a declared utility, but the area alone leaves shape unresolved. The two area metrics average over different axes: $\mathrm{mG\text{-}Pass@}k$ averages over high correctness thresholds at a fixed budget, while $\mathrm{AUC@}K$ averages over discovery budgets at threshold one. Both can identify only the weighted area, leaving the slope, endpoints, and local reversals unresolved. This is the same projection problem as in Appendix K.3, but the hidden weight vector is now part of the metric definition.

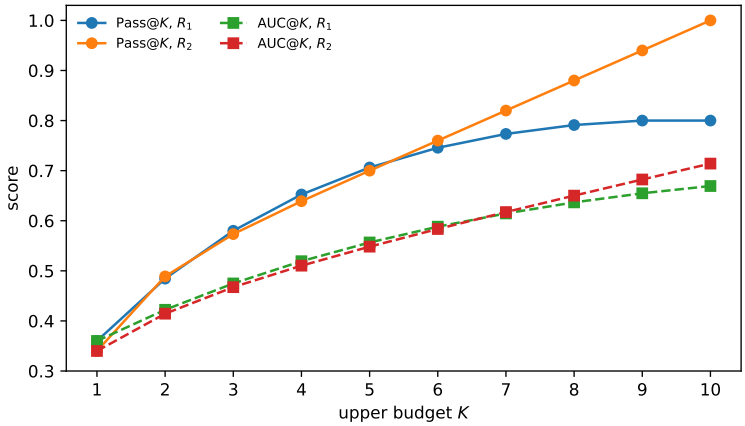

Figure 42: AUC@K as a budget-mixture projection. The solid curves are the finite-bank pass@K values for the two rank-reversal banks in Equation (47); dashed curves are the cumulative uniform budget averages. The cumulative area can preserve an earlier ranking after the operating-budget pass@K ranking has already reversed.

For mG-Pass@k, Equation (120) gives the key diagnostic. A model with $p_i = 0.49$ on every question has large-budget mG-Pass@k equal to zero despite having average accuracy 0.49. A model that is perfect on 25% of questions and impossible on the rest has average accuracy 0.25 but large-budget mG-Pass@k equal to 0.25. The latter model wins the high-threshold area because mG-Pass@k encodes a stability utility; its average correctness is lower.

Table 60: AUC@K can disagree with the operating budget. The two finite banks are the rank-reversal pair in Equation (47). At $K = 6$, pass@6 ranks R_2 above R_1, but the uniform cumulative area AUC@6 still ranks R_1 above R_2 because earlier budgets retain positive weight.

K	pass@K R_1	pass@K R_2	AUC@K R_1	AUC@K R_2
1	0.360	0.340	0.360	0.340
2	0.484	0.489	0.422	0.414
5	0.706	0.700	0.557	0.548
6	0.746	0.760	0.588	0.584
7	0.773	0.820	0.615	0.617
10	0.800	1.000	0.669	0.714

AUC@K has an analogous budget-mixture failure. Table 60 uses the finite-bank pair from Equation (47). At $K = 6$, the operating-budget discovery score ranks R_2 above R_1, while AUC@6 ranks R_1 above R_2 because the earlier budgets $1{:}5$ remain in the uniform average. At $K = 7$, the cumulative area ranking changes. The AUC score therefore depends on the upper budget and on the budget grid. Uniform weighting over $1, 2, \ldots, K$ is only one possible utility over compute; a logarithmic grid, a deployment-specific budget distribution, or a hard operating budget can induce a different ordering.

Extended test-time scaling metrics are interpretable only when their implicit operating point is made explicit. For G-Pass@k_τ and Maj@k, the report should include the threshold convention, the full threshold path or enough nearby thresholds to reveal slope, and uncertainty for the selected coordinate. For mG-Pass@k, the report should state the threshold interval and integration weights and should include endpoint values such as majority and unanimity. For AUC@K, the report should state the budget grid, the upper budget, and the operating-budget endpoint. For modal-answer majority, the response record must include answer clusters or a reproducible canonicalization rule together with binary correctness.

The common object remains the discovery–stability profile. Scalar metrics such as G-Pass@k_τ, Maj@k, mG-Pass@k, and AUC@K are useful views when the view matches the intended deployment utility. They fail as complete summaries when the threshold, majority rule, or budget weights are treated as neutral.