ARE AGENT RELIABILITY RANKINGS REAL? A STATISTICAL TREATMENT OF THE PASS^k METRIC

Anonymous authors
Paper under double-blind review

ABSTRACT

The reliability of LLM agents—their tendency to succeed *repeatably*, not merely once—is increasingly measured by pass^k, the probability that an agent succeeds on all k independent attempts of a task. It is reported by τ-bench, τ^2-bench and agent-memory benchmarks, and used to rank models. We show that its statistics have never been checked, and that its failures share one mechanism: *mis-centering*.

The ubiquitous plug-in estimator $(c/n)^k$ is provably upward biased. We give its exact bias in closed form via Stirling numbers; the *relative* bias grows without bound in k (at p=0.75, n=10: 3% at k=2, 101% at k=8). That bias then propagates silently into interval estimates. The hierarchical (task $\times$ trial) bootstrap resamples within-task counts as $\mathrm{Binomial}(n, \hat{p})$, which makes even the *unbiased* estimator unbiased for $\hat{p}^k$ rather than p^k: its resampling distribution is centred exactly on the plug-in. Coverage therefore *falls as tasks are added*—91% $\to$ 36% $\to$ 7% at M=25, 115, 230 (k=8, n=10)—which no variance-based account predicts, and re-centring does not repair. We report equally that the plain *task-level* bootstrap does *not* fail.

We then give the first pass^k interval with a finite-sample, distribution-free coverage guarantee, via betting concentration on the bounded unbiased estimator, and compare it against a hierarchical Beta–Binomial baseline. That baseline is narrower where its Beta shape holds; but on a shape agent suites actually have—an atom at p=1, tasks every agent solves—it loses coverage *as tasks are added* (0.59 at M=230), the same mis-centering signature, this time from shrinkage rather than resampling. And at the budgets suites run, neither it nor the bootstrap reaches nominal (0.90–0.93 at $M \leq 50$), where the guaranteed interval covers 1.00. Inverting the closed-form variance yields a design rule: separation is *task-limited*, not trial-limited.

Finally we test the theory on data. On 4,600 agent rollouts we run ourselves (pre-registered), Theorem 1 predicts the observed bias with no free parameters (correlation 0.969), and the plug-in overstates pass^8 by 20–147%—worst for the *least* reliable agent, where deployment is actually in question. A k=1 audit of six leaderboards—the regime *most* favourable to separation, and therefore a lower bound on the problem—finds 33% of all model pairs statistically indistinguishable and 16 of 30 models sharing τ^2-bench's top tier. We release a drop-in tool.

1 INTRODUCTION

As LLM agents move toward deployment, the binding constraint has shifted from *capability* to *reliability*: an agent that solves a task once but fails it on the next identical attempt cannot be trusted in production. This has made pass^k—the probability an agent succeeds on all k i.i.d. attempts of a task, averaged over tasks—the metric of choice for reliability, popularized by τ-bench (Yao et al., 2024) and now standard across tool-use and agent-memory benchmarks. Unlike pass@k (a *capability*/best-of-k metric with a well-studied unbiased estimator (Chen et al., 2021)), pass^k is a nonlinear, small-probability functional whose statistics have not been examined.

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.

We show this omission is consequential. pass^k is used at small samples—benchmarks typically run $k \leq 8$ with a handful of trials per task—precisely the regime where nonlinear small-probability estimation is most fragile. We make four contributions.

1. **Exact bias (§2).** The plug-in estimator $(c/n)^k$ is upward biased for p^k. We give the exact bias in closed form and a $\mathcal{O}(1/n)$ leading term, and validate both against Monte Carlo; at k=8, n=10 it overstates reliability by 8.8 points averaged over tasks (10.1 for a single p=0.75 task).
2. **One mechanism behind the interval failures, and one interval that cannot fail (§3).** The hierarchical (task × trial) bootstrap and the normal "error-bar" interval (Miller, 2024) both fail because they are centred on the plug-in rather than on θ, so their coverage *degrades as tasks are added* (36% and 19% at M=115; 7% and 2% at M=230; k=8, n=10). A hierarchical Beta–Binomial—narrower than ours where its model holds—fails the same way via shrinkage rather than resampling: under an atom at p=1 its bias is fixed in M and coverage falls to 0.59 at M=230. One mechanism, three instances. The plain task-level bootstrap does *not* fail, and we say so. We then give the first pass^k interval with a *finite-sample, distribution-free coverage guarantee*, via a betting construction (Waudby-Smith and Ramdas, 2024) on the bounded unbiased estimator, together with a permutation-symmetric variant.
3. **Sample-size requirements (§4).** Inverting the closed-form variance yields the trials/task and tasks needed to certify a reliability gap; we show comparisons are task-limited, not trial-limited, past a point.
4. **Leaderboard separability at k=1, a lower bound (§5).** Re-analyzing six real agent benchmarks (30 models each) with proper paired intervals, 33% of *all* model pairs are statistically indistinguishable and 16 of 30 models share τ^2-bench's top tier—in the regime *most* favourable to separation, so a bound on the problem rather than a measurement of $k > 1$ ranking.

Our aim is constructive: we release estimators, a guaranteed interval, and a required-N rule as a drop-in tool so the community can report pass^k correctly.

2 The pass^k estimand and its bias

Fix an agent. Task i has an unknown per-attempt success probability p_i; we observe c_i successes in n_i i.i.d. attempts, $c_i \sim \mathrm{Binomial}(n_i, p_i)$. Per task, pass$_i^k = p_i^k$; the reported aggregate is $\theta = \mathbb{E}_i[p_i^k]$ over a task distribution (estimated by the mean over M sampled tasks).

Is pass^k the right estimand? We take $\theta = \mathbb{E}_i[p_i^k]$ as given because it is the quantity benchmarks report, but it is only one of several ways to formalize "repeatably reliable." A per-task threshold $\Pr_i[p_i \geq \tau]$ (fraction of tasks the agent is at least τ-reliable on) or a Bayesian posterior over p_i (Hariri et al., 2026) encode different, equally defensible targets. We stress that our results are not specific to the "all k succeed" choice: the bias analysis, the U-statistic variance, and the distribution-free interval below apply verbatim to *any* aggregate $\mathbb{E}_i[g(p_i)]$ admitting a bounded, per-task unbiased estimator. pass^k is simply the instance the field has adopted, and the one whose statistics are currently mis-handled.

Two estimators. The *plug-in* estimator is $(c_i/n_i)^k$. The *unbiased* estimator is the falling-factorial ratio

$$\hat{\theta}_i = \binom{c_i}{k} \Big/ \binom{n_i}{k}, \qquad \mathbb{E}[\hat{\theta}_i] = p_i^k, \tag{1}$$

i.e. the probability that a uniformly random size-k subset of the n_i attempts are all successes. (This estimator is not new: it is the "all k succeed" member of the same U-statistic family as Chen et al. (2021)'s pass@k estimator $1 - \binom{n-c}{k}/\binom{n}{k}$, obtained by taking the other tail, and τ-bench already uses it unnamed. We claim no credit for it. What was missing—and what this paper supplies—is the exact bias of the plug-in that practitioners use *instead* of it, the variance of the estimator itself, and any interval with a coverage guarantee.)

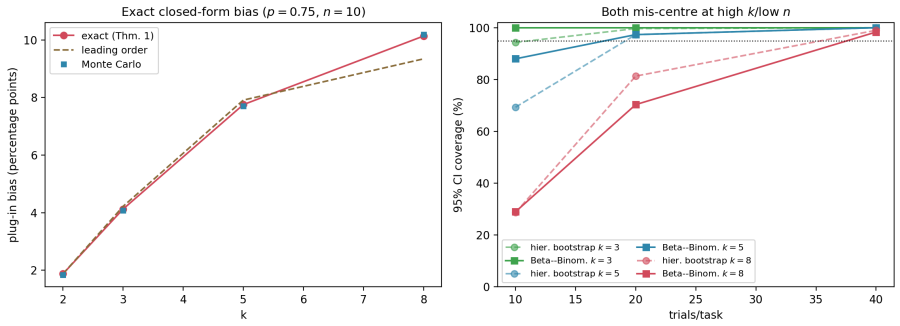

Figure 1: **Left:** exact closed-form bias (Thm. 1) matches Monte Carlo; the plug-in overstates reliability, growing with k over this range (the dependence is non-monotone in general; see §2). **Right:** 95% CI coverage vs. trials/task. The *hierarchical* (task × trial) bootstrap collapses at high k/low n for the reason given in Prop. 1; the per-task Beta–Binomial credible interval helps in the moderate regime (k=5) but not the extreme one, where both sit near 29%. The task-level bootstrap does not collapse (Table 1), nor does the hierarchical Bayesian model discussed in §3.

Theorem 1 (Exact bias of the plug-in)**.** *For* $c \sim \mathrm{Binomial}(n, p)$,

$$\mathbb{E}\big[(c/n)^k\big] \;=\; n^{-k} \sum_{j=1}^{k} S(k,j)\,(n)_j\,p^j, \qquad \text{Bias} \;=\; \frac{k(k-1)}{2}\,\frac{p^{k-1}(1-p)}{n} + \mathcal{O}(n^{-2}) \;>\; 0, \tag{2}$$

where $S(k,j)$ *are Stirling numbers of the second kind and* $(n)_j$ *the falling factorial.*

The plug-in is thus *provably upward biased* and vanishes only as $1/n$. Its dependence on k needs care, because the two readings differ. *Relative* bias, $\frac{k(k-1)}{2}\frac{1-p}{np}$, is quadratic and unbounded (3%, 101%, 897% at k=2, 8, 16 for p=0.75, n=10). *Absolute* bias is non-monotone—p^{k-1} eventually beats $k(k-1)$—and tends to p^n, not 0: the plug-in's absolute error asymptotes to the probability of a perfect run. The two can point opposite ways inside the range benchmarks use (at p=0.4 absolute bias peaks at k=3 while relative bias keeps rising), so we state which is meant throughout; Appendix A gives both, and notes that for $k > n$ no unbiased estimator exists at all—a third reason to report only $k \le n$. Reliability is read relatively, which is the operative sense and the one §5 measures, where overstatement reaches 147% for the *least* reliable model. §5 shows that this closed form predicts the bias measured on 4,600 real agent rollouts, with no fitted quantity. Figure 1(left) confirms the exact formula against 2×10^5-sample Monte Carlo for a single task at p=0.75, n=10: the closed form matches to $<10^{-3}$, and there the plug-in overstates pass^8 by 10.1 points. Averaged over the task distribution the field actually reports—$\theta = \mathbb{E}_i[p_i^k]$ with $p_i \sim \mathrm{Beta}(6, 2)$ (mean 0.75), n=10—the plug-in still overstates pass^8 by 8.8 points, while the unbiased estimator is within 0.1.

3 Confidence Intervals for Pass^k

3.1 Why Standard Intervals Fail

The aggregate variance has two components: within-task (estimating p_i^k from n_i attempts) and between-task (finite M). We give the exact within-task variance of the unbiased estimator as a U-statistic (kernel = "all k succeed"):

$$\mathrm{Var}(\hat{\theta}_i) \;=\; \binom{n}{k}^{-1} \sum_{d=1}^{k} \binom{k}{d}\binom{n-k}{k-d}\left(p^{2k-d} - p^{2k}\right), \tag{3}$$

which matches Monte Carlo (Appendix). One might expect the failures below to be failures of *width*—intervals too narrow for a skewed, near-zero, discrete statistic. They are not. They are failures of *centre*, and they are inherited directly from §2.

Proposition 1 (The hierarchical bootstrap re-injects the plug-in bias)**.** *Let $\hat{p}_i = c_i/n$ and let the within-task resample be $c_i^\star \sim \mathrm{Binomial}(n, \hat{p}_i)$, as in the standard hierarchical (task × trial) bootstrap. Then for the unbiased estimator,*

$$\mathbb{E}^\star\left[\binom{c_i^\star}{k}\Big/\binom{n}{k} \;\middle|\; c_i\right] \;=\; \hat{p}_i^{\,k},$$

so the bootstrap distribution of $\hat{\theta}$ is centred on the plug-in $\frac{1}{M}\sum_i \hat{p}_i^k$*, not on $\hat{\theta}$. By Theorem 1 the displacement is the plug-in bias: $\mathcal{O}(1/n)$, and independent of M, while the interval half-width is $\mathcal{O}(M^{-1/2})$.*

The proposition is immediate—the unbiased estimator is unbiased for the success probability of whatever binomial it is fed, and the resample feeds it $\hat{p}_i$—but its consequence is not. We should say plainly that we cannot cite a published benchmark running this variant on pass^k; we present it as a warning about a natural implementation rather than an indictment of current practice, and note that it is the one our own toolkit used before we analysed it. A fixed offset compared against a shrinking width means coverage *decreases* as tasks are *added*, which no variance-based account of the failure predicts. Table 1 confirms it: at k=8, n=10 the hierarchical bootstrap covers 91% at M=25 and 7% at M=230. Measured directly, its centre sits 0.0884 above $\hat{\theta}$, and the plug-in sits 0.0884 above $\hat{\theta}$—the same displacement to four decimals, and the same 8.8 points quoted in §2 (`code/hier_mechanism.py`). Re-centring does not repair it: the basic (reverse-percentile) interval simply moves the miss to the other side (31%). The normal "add error bars" interval on the plug-in (Miller, 2024) fails identically and for the identical reason (19% at M=115, 2% at M=230).

What does *not* fail. The plain *task-level* percentile bootstrap—resampling tasks only, treating each task's unbiased estimate $\hat{\theta}_i$ as one observation—covers at 92–95% throughout, as does the normal interval on that estimator, because $\hat{\theta}_i$ already carries its own within-task uncertainty into the across-task spread. We state this prominently for two reasons: it is what a reader should use if they do not adopt ours, and it is what our own leaderboard audit uses in §5, which would otherwise be self-undermining. The claim of this section is therefore narrower and sharper than "the bootstrap fails": *do not resample trials within tasks when estimating pass^k*, because doing so recreates exactly the bias the unbiased estimator exists to remove.

A per-task Bayesian Beta–Binomial (Jeffreys) credible interval fails for a different and more familiar reason. It improves the *moderate* regime (k=5, n=10: 69% → 88%) but cannot rescue the extreme one, because the data carry almost no information about pass^8 when only tasks with $\geq 8/10$ successes contribute.

The strongest competitor, and the shape that breaks it. The natural Bayesian competitor is not that per-task interval but a *hierarchical* Beta–Binomial—$p_i \sim \mathrm{Beta}(\alpha, \beta)$ with (α, β) fitted from the data (Hariri et al., 2026)—which pools across tasks and returns estimator and interval together. We built it (`code/hier_bayes.py`) and report the comparison in full, including where it beats us. When its model is right it is *narrower* than ours at every M (0.117 vs. 0.176 at k=8, n=10, M=115), and it is robust to the misspecification one first reaches for: under a bimodal Beta mixture its coverage holds at 0.92–0.96, because shrinkage moves the point estimate while the interval still comes from a nonparametric task bootstrap.

It breaks on a shape agent suites actually have. Let a fraction of tasks be solved by every agent on every attempt—an atom at p=1—with the remainder piled near 0. No Beta density has an atom, so empirical Bayes must smear the spike across a continuous prior and shrink exactly the tasks that carry all of pass^k's mass. The resulting point bias is −0.047 to −0.053 and, critically, *does not shrink with M*: it is the same fixed-offset-against-narrowing-width signature as Proposition 1, and it produces the same diagnostic. Coverage falls as tasks are *added*—0.88, 0.88, 0.84, **0.59** at M=25, 50, 115, 230 (Table 1). Our interval holds at $\geq$ 0.977 throughout, and so, notably, does the task-level bootstrap (0.94–0.96): what fails is not asymptotics but the imposition of a *shape*. Every failure in this paper is a mis-centering failure, and this is the third.

That is what the guarantee buys, stated exactly. Where the task-difficulty model is right and M is large, the hierarchical interval is narrower and you should use it. Where the model may be wrong—and no benchmark reports its task-difficulty distribution—it can lose a third of its nominal coverage

95% CI coverage at k=8, n=10	M=25	M=50	M=115	M=230
Hierarchical (task × trial) bootstrap	0.91	0.66	0.36	**0.07**
Normal "error bars" on the plug-in	0.81	0.55	0.19	**0.02**
Hierarchical Beta–Binomial, model correct	0.90	0.93	0.96	0.92
Hierarchical Beta–Binomial, atom at p=1	0.88	0.88	0.84	**0.59**
Task-level bootstrap (unbiased estimator)	0.93	0.92	0.94	0.95
Normal interval (unbiased estimator)	0.92	0.92	0.95	0.95
Betting interval (§3; *guaranteed*)	1.00	1.00	1.00	**0.96**

Table 1: Coverage against a 95% target; 200 replicates per cell (Monte-Carlo s.e. ≈0.015), $p_i \sim \mathrm{Beta}(6, 2)$. Every method carrying a fixed centring offset degrades as tasks are *added*, whatever produces it—resampling onto the plug-in (rows 1–2) or shrinkage onto a prior with no atom (row 4): the offset is constant in M while the width is not (Prop. 1). Methods centred on the unbiased estimator are stable. Only the betting interval is valid by construction rather than asymptotically. Reproducible via `code/bootstrap_variants.py`.

while looking tighter. Where M is small, neither asymptotic method reaches the stated level at all: 0.903 (hierarchical) and 0.910 (task bootstrap) at M=25, against a 95% target. Only the distribution-free construction is level-correct in all three regimes, at 1.4–2× the width.

3.2 A DISTRIBUTION-FREE INTERVAL WITH A GUARANTEE

Our key observation is simple and consequential: $\hat{\theta}_i \in [0, 1]$, the $\hat{\theta}_i$ are independent across tasks, and $\mathbb{E}[\hat{\theta}_i] = \theta$. Therefore *any* distribution-free confidence interval for the mean of bounded random variables applies directly and yields a *finite-sample coverage guarantee for pass^k*, for any k, n, and task-difficulty distribution. We instantiate three: Hoeffding, empirical-Bernstein (Maurer and Pontil, 2009), and the tight betting/test-martingale interval of Waudby-Smith and Ramdas (2024).

Proposition 2 (Guaranteed coverage)**.** *Let $\hat{\theta}_1, \ldots, \hat{\theta}_M \in [0, 1]$ be independent with common mean θ. The Hoeffding, empirical-Bernstein, and betting intervals each cover θ with probability at least $1 - \alpha$, with no assumption on k, n, or the distribution of p_i.*

Table 1 validates this at n=10, and Fig. 4 (Appendix E) sweeps trials/task: all three cover ≥ 95% everywhere, including k=8, n=10, where the hierarchical bootstrap gives 26–36%. Hoeffding and empirical-Bernstein are valid but loose; the betting interval is the sweet spot—at k=8, n=10 it is 31% narrower than Hoeffding (0.174 vs. 0.253), and only 15% wider than the hierarchical bootstrap (0.152) that covers a third of the time. Against the interval that *does* cover—the task-level bootstrap (0.126 at M=115)—the guarantee costs 40% width, which is the honest price of finite-sample validity rather than asymptotic adequacy.[1] To our knowledge this is the first confidence interval for pass^k with finite-sample validity.

The guarantee assumes the per-task estimators $\hat{\theta}_i$ are independent. Suites built from clustered task families (shared templates, shared underlying state) violate this; there the interval should be applied over independent task *clusters*—a grouped/blocked betting variant—which shrinks the effective sample and *widens* the interval. Our independent-task intervals are therefore, if anything, optimistic on clustered suites: the indistinguishability conclusions of §5 only strengthen under clustering, though the design rule of §4 moves the other way (§7).

Order-dependence, and a permutation-symmetric interval. The betting interval is a sequential construction, so it depends on the order in which tasks enter the capital process; on real data with systematic level shifts between task groups (e.g. domains) the tasks must be permuted first. Fixing a seed and reporting the spread is unsatisfying—it reports a *family* of intervals and distinguishes no member of it. We therefore also give a permutation-symmetric interval that keeps the guarantee. For a candidate mean m, let $K_M^{(b)}$ be the final capital under permutation b; each is an

[1] Two one-sided capital processes are run per candidate mean; Ville's inequality bounds each separately, so each must be thresholded at $2/\alpha$ for overall level α. Thresholding both at $1/\alpha$ instead yields a level-2α set—invisible at these task counts, but material at large M.

e-value under $H_0 : \theta = m$, and e-values are closed under averaging, so $\bar{E} = \frac{1}{B}\sum_b K_M^{(b)}$ satisfies $\Pr(\bar{E} \geq 1/\alpha) \leq \alpha$ by Markov's inequality. Inverting over m leaves no seed to pre-register and no spread to report. The trade is scope: $\bar{E}$ is valid at the final task count M, whereas the single-order construction is a genuine confidence *sequence*, covering simultaneously at every $t \leq M$ by Ville's inequality. In the hardest cell the symmetric variant costs $\approx 8\%$ width (0.188 vs. 0.174) at coverage 1.00 (`code/avg_eprocess.py`). Both are released; we report the single-order interval below, and the symmetric one wherever ordering could be contested.

4 How many trials and tasks do you need?

Inverting Eq. (3) plus the between-task term gives an analytic standard error and hence a required-N rule. Two findings: (i) a trustworthy 95% CI for pass^8 needs ~ 40 trials/task, whereas benchmarks run ≤ 8; (ii) to *distinguish* two agents whose true pass^8 differ by a realistic 7 points on a fixed 115-task suite, 80% power is *unreachable at any number of trials*—between-task variance dominates, so the lever is *more tasks*, not more trials. This matches an empirical power plateau at 77% and gives benchmark builders a concrete (tasks $\times$ trials) design rule. The specific plateau depends on the between-task variance estimate and assumes tasks are i.i.d.: suites built from shared templates (clustered task families) have a smaller *effective* number of tasks, which lowers achievable power further, so 77% is if anything optimistic. The qualitative conclusion—that separation is task-limited, not trial-limited—is robust to this estimate. Table 3 makes it quantitative: holding the reward gap at its true 7.2-point pass^8 value and sweeping the number of tasks M, the power *ceiling* as $n \to \infty$ (where within-task noise vanishes and only between-task variance survives) is 72% at τ-bench's M=115—below 80% at *any* number of trials, consistent with the empirical plateau—whereas *doubling the tasks* to M=230 lifts it to 94%. Trials cannot buy what tasks must.

The rule, in closed form. Let d_i be the paired per-task difference between two agents and $\sigma_d^2 = \mathrm{Var}_i(d_i)$ its between-task variance, which §3 lets us estimate directly. As $n \to \infty$ the within-task term vanishes and $\mathrm{SE}(\bar{d}) \to \sigma_d/\sqrt{M}$, so certifying a true gap δ at level α with power $1-\beta$ requires

$$M \;\geq\; \frac{(z_{\alpha/2} + z_\beta)^2\, \sigma_d^2}{\delta^2}, \qquad \text{and at finite } n, \quad M \;\geq\; \frac{(z_{\alpha/2} + z_\beta)^2 \big(\sigma_d^2 + \bar{v}(n)\big)}{\delta^2}, \tag{4}$$

where $\bar{v}(n) = \mathbb{E}_i[\mathrm{Var}(\hat{\theta}_i \mid p_i)]$ is the mean within-task variance of Eq. 3. The first bound contains no n: it is a floor no trial budget can lower. Equivalently, if $M\delta^2/(z_{\alpha/2}+z_\beta)^2 \leq \sigma_d^2$ then *no* number of trials attains the target power—the task-limited condition, stated as a checkable inequality.

Worked example, on measured quantities. From our own four-model rollouts (§5; n=10, 115 τ-bench tasks), the median paired between-task variance is remarkably flat in k—$\hat{\sigma}_d^2 = 0.0584, 0.0510, 0.0540, 0.0580$ at k=1, 2, 4, 8—while the median observed gap collapses, $0.1246 \to 0.1176 \to 0.0800 \to 0.0434$. Substituting into Eq. 4 at 80% power ($z_{\alpha/2}+z_\beta = 2.8016$), the tasks needed to certify the gap that actually exists rise from **30** at k=1 to **242** at k=8; we print four digits so the arithmetic reproduces. τ-bench has 115. It is sufficient to rank these agents on accuracy and roughly 2$\times$ short of ranking them on pass^8—and the shortfall is in tasks, since σ_d^2, the quantity that sets the floor, barely moves with k.

5 Leaderboard separability at k=1: a lower bound on the problem

We re-analyze six real agent benchmarks from released trajectories (30 frontier models each, evaluated on shared tasks), recomputing each leaderboard with a paired (shared-task) bootstrap and 95% intervals. A pair of models is *indistinguishable* if the 95% CI of their paired difference includes 0. Intervals are nominal 95% per pair and we apply no multiplicity correction across the $\binom{30}{2} = 435$ pairs of a leaderboard; since correction would widen intervals and so *raise* every indistinguishability rate we report, the uncorrected figures are the conservative choice for our claim. We write $\Pr(\text{\#1 truly best})$ for the fraction of paired bootstrap resamples in which the leaderboard's reported top model attains the maximum; it is a bootstrap approximation to a posterior probability rather than a coverage statement, and like any top-1 quantity it carries a winner's-curse optimism that makes

it, if anything, generous to the published ranking. Two scope notes. The interval used here is the *task-level* paired bootstrap, which §3 shows is the variant that does not fail; nothing in this section depends on the hierarchical bootstrap indicted there. And three of the six suites (ACEBench, Nexus-Bench, complex-func-bench) are function-calling *accuracy* benchmarks that do not themselves report pass^k; we include them because the ranking question is identical at $k{=}1$ and they supply the large-M end of the task-count range, but the pass^k-specific results of §2–3 are not tested by them.

Table 2 (Appendix B) and Fig. 2 report the result (Table 2 carries the same information in numbers). We lead with the two statistics that do not move with how finely a leaderboard is sliced: across six benchmarks 33% of *all* model pairs are statistically indistinguishable, and on τ^2-bench 16 **of** 30 **models' intervals overlap the top model's**—half the leaderboard is a single statistical tier. The reported #1 model is truly best only 68% of the time. The more quotable 91% figure for *adjacent* pairs is reported for completeness and discounted below.

Which statistics to trust. The adjacent-pair rate is the most quotable and the least robust—it falls mechanically if a shorter model list is published—so we rest the argument on the two that do not move with list length: 33% of all pairs indistinguishable, and 16 of 30 sharing the top tier. We likewise do not lead with $\Pr(\text{\#1 truly best})$: a bootstrap resampling frequency is not a calibrated posterior for an argmax over 30 candidates, and selection effects make it optimistic about the published ranking. What it shows robustly is a *relationship*—benchmarks with ≤ 273 tasks give 39–49%, those with ≥ 1000 give 82–100% (Pearson 0.97 against task count)—confirming the task-limited regime of §4 on data collected by others. Appendix B gives the per-benchmark table and the full discussion.

What the $k{=}1$ audit shows, and its scope. Public agent logs are single-attempt, so this re-analysis is at $k{=}1$, where pass^k coincides with accuracy. This is the *easiest* regime in which to separate models—the plug-in is unbiased and each per-task estimate is maximally informative—whereas at $k > 1$ the same M tasks and n trials yield a smaller, more skewed, more discrete signal, carrying the bias and coverage pathologies of §2–3 on top. As k grows the pass^k *gap* contracts while its interval widens; when one model stochastically dominates the other per task—the case our real pairs below exhibit—both forces cut separability. We do not claim this universally, and our own multi-model run below is the counterexample: two models with equal accuracy but different *consistency* (equal mean p_i, different dispersion) can separate at $k > 1$, and there $\Pr(\text{\#1 truly best})$ *rises* with k rather than falling. Such cases *reorder* rather than restore the published ranking, so the $k{=}1$ audit is best read not as a strict lower bound but as evidence that the metric is already too noisy to support these leaderboards in its most favourable regime. Where real *multi-trial* logs exist the mechanism can be measured rather than inferred, which we do next.

Real $k > 1$ evidence (a model pair). A few benchmarks release multiple attempts per task, letting us measure the $k > 1$ behavior on real data at zero additional cost. Using 1,980 released τ-bench rollouts (zero errors) for two models over 165 shared tasks (`gpt-4o`, $n{=}4$ trials/task; `claude-3.5-sonnet`, $n{=}8$), we compute the paired pass^k difference and its betting interval for $k = 1, \ldots, 4$ (Fig. 3). As k grows the paired-difference interval *widens* ($0.200 \to 0.226$, $+12\%$) and separability $|\text{gap}|/\text{width}$ *shrinks* ($0.367 \to 0.279$, -24%): the two models become progressively harder to distinguish at higher reliability thresholds, exactly as the theory predicts. Two robustness checks bound the claim. The betting interval is order-dependent, so these are one pre-registered draw; over 200 random task orderings separability falls in 100% of them (median -22%, worst case -10%) and the interval widens in 96%. The two arms were logged at different depths, and at $k{=}4$ the $n{=}4$ arm's estimator degenerates to $\not\Vdash[c{=}4]$, so part of the widening could be an artifact of the unequal budget; holding both to $n{=}4$ by subsampling the deeper arm leaves it intact (-21.7% separability, $+12.7\%$ width, against -24.0% and $+12.5\%$ unmatched). Monotonicity is the weaker half: the $k{=}1 \to 4$ endpoint effect holds in 89% of subsampling draws, but the step-by-step curve is monotone in only 52% and the 5–95% band $[-48\%, +12\%]$ crosses zero, so we claim the endpoint degradation, not a trend. With two models the leaderboard-level rates are degenerate; this measures the per-pair *mechanism* only. Table 4 in Appendix D gives the per-model curves—pass^k falls for both while the paired gap holds near 0.07, so the widening interval, not a shrinking gap, is what erases separability. The four-model run below removes the mismatched-budget caveat entirely.

A pre-registered multi-model run, and what it did and did not show. The released-trajectory pair is not ours and carries a mismatched trial budget, so we ran our own: four models on all 115 τ-bench retail tasks at n=10 trials each, $k \in \{1, 2, 4, 8\}$, 4,600 rollouts, ∼$47, with an independent user simulator and a pre-registered analysis plan, decision rules and null (Appendix E). Three results, reported in the order the evidence supports rather than the order we hoped for.

(i) Theorem 1 predicts the bias on real data. Feeding each task's own $\hat{p}_i$ into the closed form of Thm. 1 predicts the gap between the plug-in and the unbiased estimator across all twelve (model, k>1) cells with *no fitted quantity*: correlation 0.969, mean absolute error 0.0021, mean ratio predicted/observed 0.94. The overstatement is large and uneven—20%, 31%, 38% and 147% of the true value at k=8—and it is worst for the *least* reliable agent ($\bar{p} = 0.152$), which is exactly where a practitioner is deciding whether an agent is deployable at all. This is the paper's central estimator claim, measured rather than simulated.

(ii) The per-pair mechanism replicates. Separability |gap|/width falls from k=1 to k=8 in five of six pairs (Table 5), by −33% to −85% among the four pairs that are resolvable at k=1. This is the same effect as the released-trajectory pair, now on four models at a *matched* n=10, so it cannot be an artifact of unequal trial budgets. It is robust to the analysis rule: relaxing complete cases to $n_i \geq k$, which restores the task set from 68 to 115, leaves every drop within 6 points.

(iii) The leaderboard-level hypotheses returned their pre-registered null, and not for want of power. Adjacent-pair indistinguishability rose (0.33 → 0.67) but survived resampling in only 52% of draws; Kendall's τ_b between the k=1 and k=8 orderings was 0.667 with an interval touching 1.0; and Pr(#1 truly best) *rose* with k (0.83 → 0.96), refuting our own H2. The cause is the slate, and Eq. 4 quantifies it. Two gap estimates exist for this pair and they differ by an order of magnitude, so we state both and say which the rule governs. Their *unpaired* pass^1 values differ by 0.0023 (0.3989 vs. 0.3966), but those are computed over each model's own complete cells—89 tasks and 88, not the same 89 and 88—so the comparison is confounded by task composition. The quantity Eq. 4 governs, and the one the paired bootstrap actually tests, is the *paired* difference on the 77 tasks both models complete: $\hat{\delta} = 0.02727$ with $\hat{\sigma}_d^2 = 0.05438$, giving $M \geq \mathbf{574}$ tasks at 80% power ($z_{\alpha/2} + z_\beta = 2.8016$). τ-bench has 115. The run was therefore five times short of resolving its own closest pair *before it was executed*—predicted in advance by our design rule rather than discovered afterwards—and the reordering these statistics detect is two tied models exchanging places. (The unpaired figure would imply ∼80,000 tasks; it is not the estimand the rule or the test refers to.) Restoring the discarded 41% of tasks does not change it. We therefore make no leaderboard-level $k > 1$ claim; the $k > 1$ evidence is the mechanism in (ii) plus the simulation of §2–3. Appendix E gives the pre-registration, the deviation log, and one analysis defect we left standing rather than repair after the fact.

6 Related work

τ-bench introduced pass^k (Yao et al., 2024); τ^2-bench (Barres et al., 2025) and agent-memory benchmarks inherit it. The unbiased estimator is classical—the "all k succeed" member of Chen et al. (2021)'s pass@k family—but has never been paired with an analysis of its bias, variance, or a valid interval. Hariri et al. (2026) give a Bayesian framework for pass@k, the opposite (capability) estimand, and supply the hierarchical baseline we benchmark in §3; Jiang et al. (2026) document a different failure of the same family (setting n to unit tests rather than independent rollouts), fixing *what* the estimator is applied to where we characterise its statistics once applied correctly. Rabanser et al. (2026) and Raj et al. (2026) broaden what reliability means; we instead ask whether the metric already reported at scale supports the comparisons drawn from it—and at $n \leq 8$ trials, the asymptotic normality their tests assume is precisely what fails. Miller (2024) add error bars for mean accuracy, invalid for pass^k; Kapoor et al. (2024), Zhu et al. (2025) and Stroebl et al. (2025) address benchmark rigour and infrastructure; Neuhof and Benjamini (2026a) and Neuhof and Benjamini (2026b) build rank CIs for accuracy leaderboards, complementary machinery that could layer on our per-task intervals. We build on empirical-Bernstein (Maurer and Pontil, 2009) and betting intervals (Waudby-Smith and Ramdas, 2024); our contribution is that the unbiased pass^k estimator is bounded and independent across tasks, so these yield the first finite-sample-valid pass^k interval.

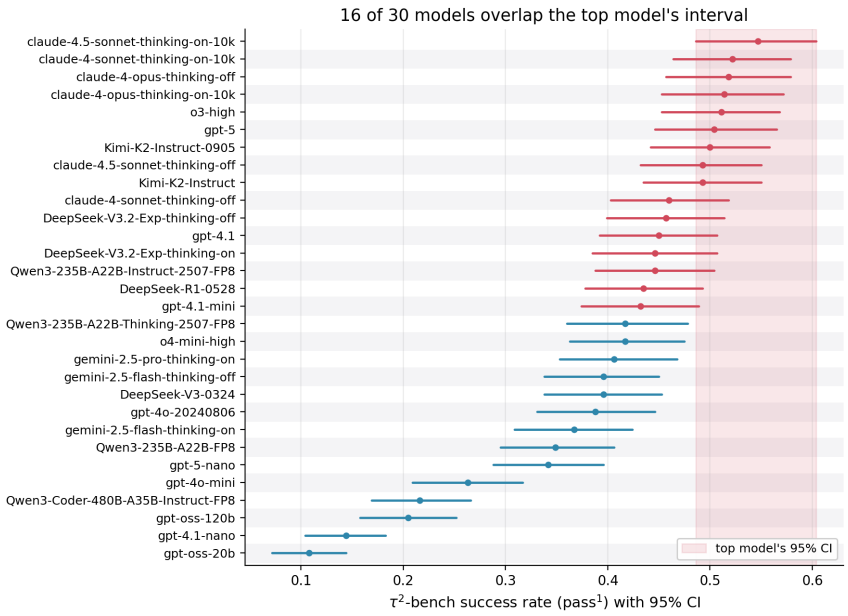

Figure 2: τ^2-bench: 30 real models on 278 shared tasks with 95% CIs. The top model's interval (shaded) overlaps 16 of 30 models; the reported ranking between neighbors is largely unsupported.

7 Limitations

The leaderboard audit is at k=1, and our $k > 1$ run was under-designed. Public agent logs are single-attempt, so §5's six-benchmark result concerns the regime where pass^k coincides with accuracy—a lower bound, not a measurement of $k > 1$ ranking, and three of the six suites do not report pass^k. Our pre-registered four-model run confirmed the estimator theory and the mechanism but returned its null on the leaderboard hypotheses, for a reason Eq. 4 states exactly: its closest pair needed 574 tasks and had 115. Inverted, it prescribes what a re-run needs: adjacent pass^8 gaps of $\delta \geq 0.061$ at M=115, hence candidates spaced $\geq$0.10 apart on published *same-domain* accuracies—our disjoint-domain calibration protected selection integrity at the cost of not transferring.

A narrower interval exists where its model holds. The hierarchical Beta–Binomial beats ours on width at every M when task difficulty is Beta-shaped, and survives smooth misspecification. Ours earns its width at small M, where no asymptotic interval reaches the stated level, and under an atom at p=1, where the hierarchical interval drops to 0.59. Since no benchmark reports its task-difficulty distribution, a user cannot generally tell which regime they are in—that uncertainty, not a width advantage, is the case for a guarantee. §3's indictment is likewise scoped: the task-level bootstrap does not fail, and we report that alongside.

Tasks are treated as i.i.d. draws. A builder treating τ-bench's 115 tasks as the inferential target rather than a sample would drop the between-task term and §4 would not apply. Clustering is likewise asymmetric: it widens intervals, strengthening §5, but makes §4's budgets *lower* bounds.

8 Conclusion

pass^k became the reliability metric of record before its statistics were understood. Its standard estimator is biased—exactly, in closed form, confirmed on 4,600 real agent rollouts—and re-enters through the hierarchical bootstrap, the plug-in normal interval, and shrinkage onto a prior with no atom, each worsening coverage as tasks are *added*. We give the first pass^k interval with a finite-sample guarantee, priced against a hierarchical Bayesian competitor, and a design rule that says when separation is unattainable at any trial budget. At k=1 a third of all model pairs across six leaderboards are indistinguishable; whether that persists at $k > 1$ remains open.

USE OF LARGE LANGUAGE MODELS

We used LLM-based coding and writing assistants in implementing the estimator and analysis code, in deriving the closed-form bias and variance expressions, in locating related work, and in drafting and editing the text. All reported numbers are independently recomputed from the raw trial logs by `code/verify.py`; the closed-form expressions are validated against Monte Carlo simulation (§2); and the $k > 1$ study was pre-registered, with every deviation recorded before the affected result was observed. The authors have checked all text, code, and derivations, including those originating with an LLM, and take full responsibility for the contents of this paper.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

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, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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 (ICLR)*, 2026. arXiv:2510.04265.

Jiajun Jiang, Sharon Zheng, Natan Vidra, and Spurthi Setty. Beyond pass@k: Measuring reliability and security of agentic code generation. *arXiv preprint arXiv:2608.14711*, 2026.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

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

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. *arXiv preprint arXiv:2607.16259*, 2026a.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. *arXiv preprint arXiv:2606.08679*, 2026b.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of AI agent reliability. In *International Conference on Machine Learning (ICML)*, 2026.

Harsh Raj, Niranjan Orkat, Suvrorup Mukherjee, Aritra Guha, Cheryl Flynn, and Subhabrata Majumdar. Consistency as a testable property: Statistical methods to evaluate AI agent reliability. *arXiv preprint arXiv:2605.10516*, 2026.

Benedikt Stroebl et al. Holistic agent leaderboard: The missing infrastructure for ai agent evaluation. *arXiv preprint arXiv:2510.11977*, 2025.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Yuxuan Zhu, Tengjun Jin, et al. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825*, 2025.

A Proof of Theorem 1 (exact bias of the plug-in)

Let $c \sim \mathrm{Binomial}(n, p)$ and write $(x)_j = x(x-1)\cdots(x-j+1)$ for the falling factorial. We use two standard identities:

1. **Raw powers in falling factorials (Stirling, 2nd kind).** For every integer c, $c^k = \sum_{j=0}^{k} S(k,j)\,(c)_j$, where $S(k,j)$ is the number of partitions of a k-set into j nonempty blocks, and $S(k,0) = 0$ for $k \geq 1$.
2. **Binomial factorial moments.** $\mathbb{E}[(c)_j] = (n)_j\, p^j$. (Each of the $(c)_j$ ordered j-tuples of distinct successful attempts occurs with probability p^j, and there are $(n)_j$ index tuples.)

Taking expectations of the first identity and applying the second,

$$\mathbb{E}\big[(c/n)^k\big] \;=\; n^{-k}\,\mathbb{E}[c^k] \;=\; n^{-k}\sum_{j=1}^{k} S(k,j)\,(n)_j\,p^j, \tag{5}$$

which is the exact expression of Theorem 1. Subtracting p^k gives the exact bias.

Leading term. Expand $(n)_j = n^j\big(1 - \frac{1}{n}\binom{j}{2} + \mathcal{O}(n^{-2})\big)$ and collect the two highest terms of (5):

$$\begin{aligned} j = k: &\quad n^{-k} S(k,k)(n)_k\, p^k = p^k\Big(1 - \tfrac{k(k-1)}{2n}\Big) + \mathcal{O}(n^{-2}),\\ j = k-1: &\quad n^{-k} S(k,k-1)(n)_{k-1}\, p^{k-1} = \tfrac{k(k-1)}{2}\,\tfrac{1}{n}\,p^{k-1} + \mathcal{O}(n^{-2}), \end{aligned}$$

using $S(k,k) = 1$ and $S(k,k-1) = \binom{k}{2} = \frac{k(k-1)}{2}$; the terms $j \leq k-2$ are $\mathcal{O}(n^{-2})$ because $n^{-k}(n)_j = \mathcal{O}(n^{j-k})$. Hence

$$\mathrm{Bias} = \mathbb{E}[(c/n)^k] - p^k = \frac{k(k-1)}{2}\,\frac{p^{k-1}(1-p)}{n} + \mathcal{O}(n^{-2}) \;>\; 0,$$

which is the leading term of Theorem 1. It vanishes at $k{=}1$ (the plug-in is unbiased for a mean) and decays as $1/n$. For $k > n$ no unbiased estimator of p^k exists at all, since $\binom{c}{k} = 0$ for every attainable c. In k the bias is non-monotone: the prefactor $\frac{k(k-1)}{2}$ grows while p^{k-1} decays, so the leading term is maximised near the solution of $\frac{1}{k} + \frac{1}{k-1} = \ln(1/p)$ (approximately $2/\ln(1/p)$ for large k; this under-states the argmax of the *exact* bias at moderate n, which is $k{=}4$ at $p{=}0.5, n{=}10$ against a leading-order 2.89). The exact bias tends to p^n, not 0, as $k \to \infty$. Relative bias $\frac{k(k-1)}{2}\,\frac{1-p}{np}$ is quadratic and unbounded. □

B The $k{=}1$ leaderboard audit in full

Which of these statistics are density-invariant. The adjacent-pair rate is the most quotable and the least robust: it depends on how finely a leaderboard is sliced, so publishing a shorter model list would lower it mechanically. On τ^2-bench the mean adjacent gap is 0.015 against a mean CI half-width of 0.056, and 27 of 29 adjacent gaps fall below a single half-width. We therefore read it as a description of leaderboards *as presented*, and rest the argument on the two statistics that do not move with list length: the fraction of all pairs indistinguishable (33%) and the number sharing the top tier (16 of 30).

$\Pr$(#1 truly best) tracks the task budget, not the benchmark. We treat this statistic with caution and do not lead with it: a bootstrap resampling frequency is not a calibrated posterior for an argmax over 30 candidates, and selection effects make it optimistic about the published ranking. What it does show robustly is a relationship. The three benchmarks with ≤ 273 tasks give 39%, 40%, 49%; the three with ≥ 1000 give 82%, 100%, 100% (Pearson 0.97 against task count)—an empirical confirmation of the task-limited regime of §4, on data collected by others. The 68% mean is therefore a property of *this* sample of benchmarks; the relationship is not. Formalising "is the reported #1 truly best" as a best-arm-identification problem (Neuhof and Benjamini, 2026a;b) would put it on firmer footing, and we leave that to future work.

Benchmark	#tasks	% adj. pairs indist.	% all pairs indist.	P(reported #1 truly best)
multi-challenge	273	100%	39%	40%
τ^2-bench	228	97%	38%	49%
τ-bench	115	93%	51%	39%
ACEBench	1023	90%	36%	82%
NexusBench	1118	90%	20%	100%
complex-func-bench	1000	76%	12%	100%
mean		**91%**	**33%**	**68%**

Table 2: Six real agent leaderboards (30 models each), re-analyzed with proper paired confidence intervals. Most adjacent leaderboard pairs are statistically indistinguishable; the reported #1 model is truly best only 68% of the time on average. The adjacent-pair column depends on how many models a leaderboard lists and should be read alongside the density-invariant "all pairs" column; Pr(#1 truly best) rises sharply with task count (Pearson 0.97), which is the task-limited prediction of §4. All six rows use one uniform cross-benchmark task-id canonicalization, so that every benchmark is counted the same way; it resolves 228 tasks for τ^2-bench. (The dedicated τ^2-bench leaderboard in Fig. 2 resolves 278 native task ids over the same 30 models and gives the same qualitative conclusion.)

tasks M	power ceiling ($n\to\infty$)	% task-draws reaching 80%	median n for 80%
58	54%	25%	40
115	72%	49%	30
230	94%	90%	15
460	99.8%	100%	10
920	100%	100%	8

Table 3: Power is task-limited. Fixed true pass^8 gap (7.2 pts; $A \sim \mathrm{Beta}(7.8, 2.2)$, $B \sim \mathrm{Beta}(7.3, 2.7)$). The $n \to \infty$ ceiling is set entirely by between-task variance: at τ-bench's 115 tasks it caps at 72%, so no number of trials reaches 80% power—the lever is more *tasks*. Reproducible via `code/sensitivity_tasks.py`.

C Power as a function of the task budget

Table 3 gives the simulation behind §4: the power ceiling as $n \to \infty$ for a fixed true pass^k gap, swept over the number of tasks M. It is one Beta pair, not a general law; the general statement is the variance decomposition in §4, of which this is an instance.

D The released-trajectory model pair

Table 4 and Figure 3 give the per-model curves for the two models whose multi-trial rollouts τ-bench released (1,980 rollouts, 165 shared tasks, zero errors), discussed in §5.

E The pre-registered four-model $k > 1$ run

Four models (`gpt-5-mini`, `gpt-4.1-mini`, `gpt-5-nano`, `gpt-4.1-nano`) on all 115 τ-bench retail tasks, n=10 trials, $k \in \{1, 2, 4, 8\}$, 4,600 rollouts, ~$47, user simulator `gpt-4o-mini` (disjoint from the agent slate). Hypotheses, decision rules, the contraction order under budget pressure, and the null were fixed before any evaluation rollout; candidates were calibrated on 20 *airline* tasks, a disjoint domain, so no evaluation task informed selection.

Declared before the run and borne out. The widest adjacent gap after freezing was 0.116, below the betting rule's ≈0.15 sensitive band, so that rule was expected to have no power at this slate; it was pinned at 1.000 for every k and is recorded as uninformative rather than confirmatory. L=4 was the pre-approved abort floor, at which Kendall's τ_b moves in steps of 0.333.

model	pass^1	pass^2	pass^3	pass^4
`gpt-4o` (n=4)	.548 [.479, .627]	.425 [.356, .516]	.367 [.298, .464]	.327 [.261, .431]
`claude-3.5-sonnet` (n=8)	.622 [.556, .692]	.500 [.434, .584]	.434 [.366, .524]	.390 [.323, .481]
paired gap $\lvert A - B \rvert$	.073	.075	.068	.063
betting CI width	.200	.205	.215	.226
separability (gap/width)	.367	.365	.313	.279

Table 4: Real multi-trial pass^k (τ-bench, 165 shared tasks; betting 95% CIs; 1,980 rollouts, 0 errors). Reliability drops with k for both models; the paired gap stays ≈ 0.07 while its interval *widens*, so separability $\lvert\text{gap}\rvert$/width falls monotonically ($0.367 \to 0.279$). Leaderboard-level rates require > 2 models and are not computed here.

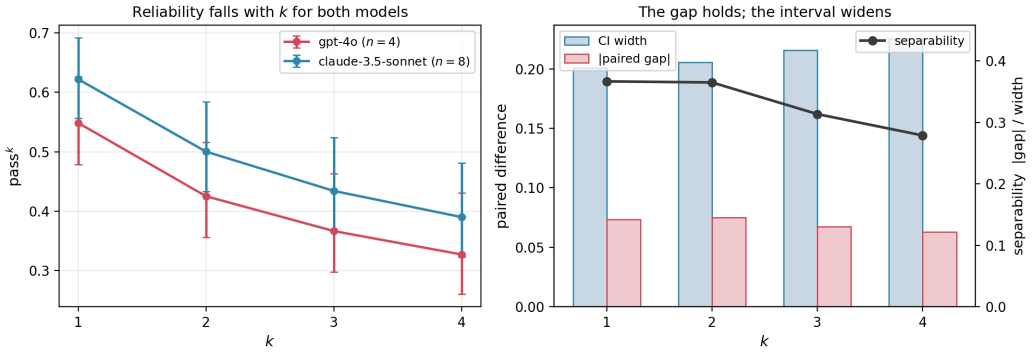

Figure 3: Real multi-trial data (τ-bench, 165 shared tasks; `gpt-4o` vs `claude-3.5-sonnet`; 1,980 rollouts, 0 errors). As k increases, the paired pass^k-difference betting interval widens (+12%) and separability $\lvert\text{gap}\rvert$/width falls (−24%), both monotonically: direct evidence that higher reliability thresholds make models harder to separate on real data, not just in simulation.

Reported results. Estimator theory validated on real data and the per-pair mechanism replicated (§5, Table 5); H1 directional but robust in only 52% of resampled draws; H2 refuted, Pr(#1 truly best) rising $0.83 \to 0.96$; H4 not established, $\tau_b = 0.667$ with an interval touching 1.0.

A budgeting consequence. Discarding tasks with incomplete trial sets—the natural complete-case rule—retains only $(1-\varepsilon)^n$ of cells at per-rollout error rate ε. In our own run $\varepsilon \approx 4\%$ at n=10 left 68 of 115 tasks common to four models: 41% of the scarce resource lost to a failure rate that looks negligible per rollout. It is avoidable—the unbiased estimator needs only $n_i \geq k$, not constant n.

One defect left standing. The trial-subsampling surface re-applies the complete-case filter at each n, so the task population shifts with n and the apparent matched-n effect is confounded. It is a one-line fix, but the numbers had already been observed, so repairing it and re-reporting would be precisely the post-hoc change the protocol forbids. It is recorded as a defect and excluded from every claim in this paper; the task-subsampling surface, which does not have this problem, is retained.

F THE PRE-REGISTERED THREE-MODEL $k > 1$ LEADERBOARD

We ran three models (`gpt-4o-mini`, `gpt-4.1-mini`, `gpt-4.1-nano`) on all 115 τ-bench retail tasks, n=5 trials each, $k \in \{1, 2, 3, 5\}$ (1,725 rollouts, ∼$22; API errors mapped to missing, never to task failures). The design, decision rules and stopping conditions were fixed in advance and every deviation was recorded before the affected result was observed.

Under the paired-bootstrap decision rule this paper's k=1 audit uses, the fraction of *adjacent* pairs that are indistinguishable rises monotonically—50% at k=1 to 100% by k=2 (Table 6)—and paired separability $\lvert\text{gap}\rvert$/width falls monotonically across all three pairs, reproducing the two-

model pair	k=1	k=2	k=4	k=8	change	resolvable at k=1
gpt-4.1-mini / gpt-4.1-nano	1.630	1.172	0.905	0.729	-55%	yes
gpt-4.1-nano / gpt-5-mini	1.530	1.140	0.740	0.507	-67%	yes
gpt-5-mini / gpt-5-nano	1.457	1.387	0.906	0.498	-66%	yes
gpt-4.1-mini / gpt-5-nano	1.062	1.117	0.901	0.712	-33%	yes
gpt-4.1-nano / gpt-5-nano	0.617	0.245	0.051	0.094	-85%	yes
gpt-4.1-mini / gpt-5-mini	0.250	0.028	0.264	0.302	$+21\%$	**no**

Table 5: Paired separability $|\text{gap}|/\text{width}$ at each k, all six pairs from the pre-registered four-model run (115 tasks, n=10, task-level paired bootstrap). Separability falls in every pair whose difference is resolvable at k=1. The last row is the tied pair (0.0023 apart in accuracy, indistinguishable at every k), where $|\text{gap}|/\text{width}$ is a ratio of noise to noise and carries no information; it is shown rather than dropped because it is also the reason the leaderboard-level hypotheses could not move.

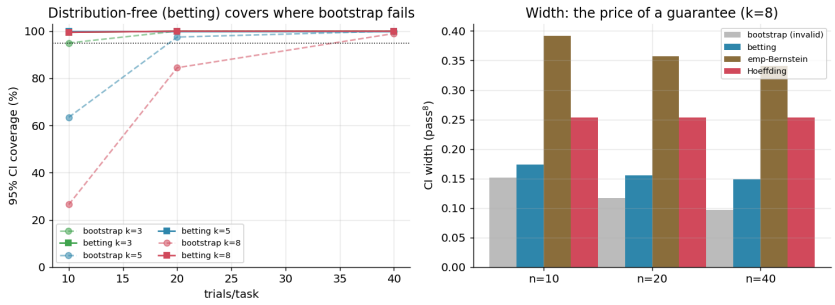

Figure 4: **Left:** coverage vs. trials/task. The distribution-free betting interval (solid) holds $\geq 95\%$ everywhere; the *hierarchical* bootstrap (dashed) collapses at low n for $k = 5, 8$, for the reason given in Prop. 1. **Right:** interval width at k=8. The betting interval is competitive with (and at n=10 tighter than) that invalid bootstrap; Hoeffding/empirical-Bernstein are valid but wider—the price of a guarantee. Table 1 gives the comparison against the task-level bootstrap, which does not collapse.

model mechanism at L=3; the reported #1 also changes with k (`gpt-4o-mini` through k=3, `gpt-4.1-mini` at k=5).

We present this as a supporting *existence proof*, not a quantitative leaderboard, and state its limits plainly. Absolute reliability is low ($\text{pass}^1 \approx 0.11$–$0.17$, far below τ-bench's published $\sim$0.6): these are small models against a lightweight user simulator that also serves as one of the three agents, following τ-bench's own gpt-4o-as-both precedent. Only $k \leq 5$ is reached, and with L=3 the adjacent-pair rate moves in coarse steps of 0.5. Two pre-registered checks temper the reading: the conservative *betting* rule is uninformative at this (L, T) (pinned at 100% for every k), and $\Pr(\#1 \text{ truly best})$ is *non-monotone* $(0.61, 0.84, 0.73, 0.56)$, so the "top is truly best falls with k" trend is not cleanly confirmed on this run. The controlled simulation (§2–3) and the two-model measurement remain the primary $k > 1$ evidence; this run adds that the leaderboard-level effect is *visible on real multi-model data* under the field's own decision rule.

G Unbiasedness and variance of the falling-factorial estimator (Eq. 3)

Let $\hat{\theta} = \binom{c}{k} / \binom{n}{k}$. It is the probability that a uniformly random size-k subset S of the n attempts consists entirely of successes, i.e. a U-statistic with kernel $h(S) = \prod_{a \in S} \mathbf{1}[\text{attempt } a \text{ succeeds}]$.

Unbiasedness. Attempts are i.i.d. Bernoulli(p), so $\mathbb{E}[h(S)] = p^k$ for any fixed S, and averaging over S gives $\mathbb{E}[\hat{\theta}] = p^k$.

k	1	2	3	5
adj. pairs indistinguishable (bootstrap rule)	50%	100%	100%	100%
adj. pairs indistinguishable (betting rule)	100%	100%	100%	100%
$\Pr$(reported #1 truly best)	0.61	0.84	0.73	0.56
reported #1 model	4o-mini	4o-mini	4o-mini	4.1-mini

Table 6: Pre-registered three-model $k > 1$ leaderboard (`gpt-4o-mini`, `gpt-4.1-mini`, `gpt-4.1-nano`; τ-bench retail, 115 tasks, n=5, 1,725 rollouts). Under the paper's bootstrap rule, adjacent-pair indistinguishability reaches 100% by k=2 and the reported #1 changes at k=5. Caveats (§7): the conservative betting rule is uninformative at L=3; $\Pr$(#1 best) is non-monotone; absolute pass^k is low and $k \leq 5$. A supporting existence proof, not a quantitative leaderboard.

Variance. For two size-k subsets S, S' with $|S \cap S'| = d$, their union has $2k - d$ distinct attempts, so

$$\mathrm{Cov}\big(h(S), h(S')\big) = \mathbb{E}[h(S)h(S')] - p^{2k} = p^{2k-d} - p^{2k} =: \zeta_d.$$

Fixing S, the number of S' meeting it in exactly d indices is $\binom{k}{d}\binom{n-k}{k-d}$. Averaging over the $\binom{n}{k}$ choices of S and noting $\zeta_0 = 0$,

$$\mathrm{Var}(\hat{\theta}) = \binom{n}{k}^{-1} \sum_{d=1}^{k} \binom{k}{d}\binom{n-k}{k-d}(p^{2k-d} - p^{2k}),$$

which is Eq. (3). The d=k term ($\zeta_k = p^k - p^{2k}$) dominates at small n, giving the near-Bernoulli behavior that makes normal intervals fail in the high-k/low-n regime. □