What pass@k Measures When Attempts Share State: Count Runs, Not Attempts

Anonymous authors
Paper under double-blind review

Abstract

Search trees, planners and self-correcting agents make the attempts of one run share a random state, yet their results are often scored with a pass@k estimator built for independent draws, which public harnesses apply to whatever attempts they receive. Prior work separates the procedure's success probability from the independent-attempt one; we ask which of them such a number measures and how to report it validly, taking the complete run as the sampling unit. Whether a run succeeds within k attempts is an unbiased estimate of the procedure's success probability succ@k under any within-run dependence, and succ@k is the only quantity estimable from the first k attempts, with estimates in $[0, 1]$, that equals pass@k when attempts are independent. Pooling attempts across runs estimates a third quantity, which we give in closed form and which, for attempts conditionally i.i.d. within a run, lies between succ@k and the independent-attempt target pass@k^{iid}. Under a calibration condition and over a rich class of dependence structures, pass@k^{iid} has a uniformly unbiased estimator exactly when at least k complete runs are observed, however much is recorded inside each run, while per-question run-level confidence sequences are valid at any number of runs, for succ@k and, under the same condition, for pass@k^{iid}. In 8 designs spanning four dependence mechanisms, the pooled report exceeds the run-level estimate of succ@4 in every design, by +0.0205 to +0.1766, and on one design a run-level interval contains a fresh reference 0.952 of the time, against 0.104 for an attempt-level one. The fix is to name the target and count fresh runs.

1 Introduction

Consider a tree search that expands four candidate solutions from one sampled plan, or a self-correcting agent that revises one draft four times. When such attempts are scored with pass@k, the standard estimator of Chen et al. (2021) is unbiased, when the draws are independent, for $1-(1-p)^k$: the probability that at least one of k *independent* draws succeeds. The four attempts of one run are not independent, though. They share the plan, the prefix or the draft, so their outcomes are dependent. What does the reported number then measure, and how precisely?

The distinction is not academic. Across the eight designs of §6, pooling all attempts of a question gives a pass@4 above the run-level estimate of the procedure's own success within four attempts in every design, by +0.0205 to +0.1766 (the *pooled excess*). Under the model of §3 the pooled number also moves when more evaluation runs of an unchanged procedure are added (§3.2), and its attempt-level interval misleads: on Arm A (Qwen2.5-1.5B-Instruct continuing shared reasoning prefixes) it contains a fresh 128-run reference only 0.104 of the time, against 0.952 for a run-level Clopper–Pearson interval (Clopper and Pearson, 1934; Figure 2b). An audit of published practice is consistent with this. Of ten recent evaluations of multi-attempt methods that we audited (an illustrative set, not a prevalence estimate), six run attempts that share state and three of those report the result next to an independent-sample success rate at $k \geq 2$; one of these three applies the independent-draw estimator to its headline candidates without stating how those candidates are constructed or how many are scored. All three public scoring harnesses we inspected apply that estimator to whatever samples they receive (Appendix D.2).

We answer three questions: which success probability a dependent-attempt pass@k reports, whether the independent-attempt quantity can be estimated from the runs actually made, and which interval

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

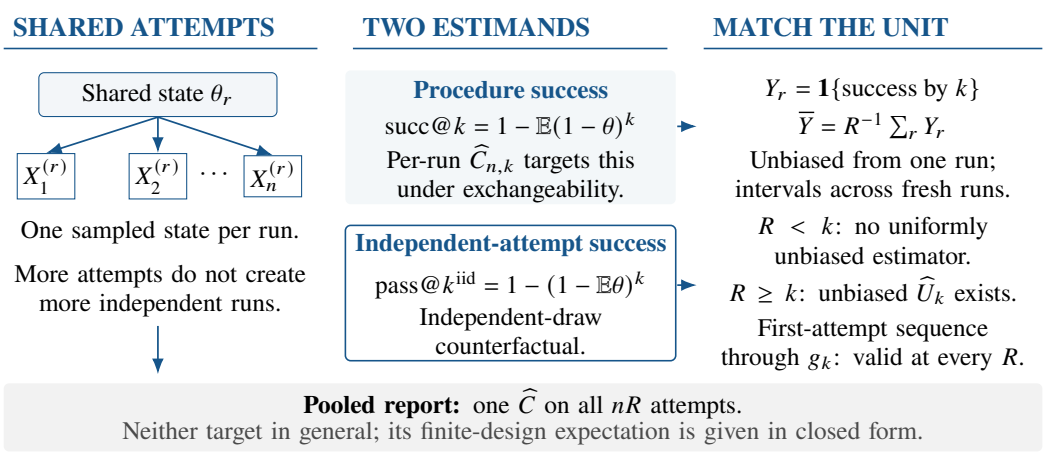


Figure 1: Attempts that share state lead to two targets, and complete runs decide which one can be estimated and how. $\mathrm{succ@}k$ is estimable from one run; uniformly unbiased estimation of $\mathrm{pass@}k^{\mathrm{iid}}$ needs $R \geq k$ runs over the rich class of §3; run-level intervals are valid at every R under Assumption 5.1 (§5.2).

is valid. The key is to take the complete run, not the attempt, as the sampling unit (Figure 1). A dependent-attempt $\mathrm{pass@}k$ can mean the probability that the procedure succeeds within k attempts, $\mathrm{succ@}k$, or the probability that k independent attempts would, $\mathrm{pass@}k^{\mathrm{iid}}$; pooling reports a third quantity, between the two when attempts are conditionally i.i.d. within a run.

Contributions.

- **What each pipeline estimates** (§3). Scoring each run by whether it succeeds within k attempts estimates $\mathrm{succ@}k$ under any within-run dependence, and $\mathrm{succ@}k$ is the only estimable extension of the independent-attempt target whose estimates stay in $[0, 1]$ (Theorem 3.5). A pooled report estimates a third, design-dependent quantity, given in closed form, that lies between the two targets for conditionally i.i.d. attempts (Proposition 3.2).
- **How many runs** $\mathrm{pass@}k^{\mathrm{iid}}$ **needs** (§4). Over the rich class of §3, a uniformly unbiased estimator exists exactly when $R \geq k$ complete runs are observed, however much is recorded inside each; one run short, the smallest achievable worst-case bias over classes containing the two-point family on $\{0, 1\}$ with weights dense in $[0, 1]$ is $2 \cdot 4^{-k}$ (Theorems 4.1 and A.6).
- **Valid intervals that count runs** (§5). Run-level confidence sequences valid at any stopping time, for $\mathrm{succ@}k$ and, under a calibrated kernel, for $\mathrm{pass@}k^{\mathrm{iid}}$; attempt-level intervals collapse; only fresh runs lower the variance floor (§F.4; Propositions A.22, A.24; Theorem A.20).
- **Evidence across four dependence mechanisms** (§6). Four shared-prefix designs, a public human-labelled reasoning tree, a self-debugging loop at two sizes and a tree-search scaffold: the pooled report exceeds the run-level estimate of $\mathrm{succ@}4$ in 8 of 8 designs with a question-bootstrap interval above zero, run-level intervals have the better interval score in 8 of 8, and the gap between the two targets replicates on five further open-weight models.

2 Related work

Targets, estimability and the run. Prior work gives the sign of the gap between the two targets and its equality case (Tran-Truong and Le, 2026, Thm. 4; Bay and Yearick, 2026, Prop. 3), names $\mathrm{succ@}k$ (Liu et al., 2025) and defines the same quantity for sequential attempts as $\mathrm{discovery@}k$ (Hübotter et al., 2026, §5, Eq. 4), gives the i.i.d. threshold $n \geq k$ for $\mathrm{pass@}k$ (Walder and Karkhanis, 2025, Cor. 2), argues for replicating complete runs (Bjarnason et al., 2026; Jiang et al., 2026; D'Oro et al., 2026; Mustahsan et al., 2025), and reads $\mathrm{pass@}k$ on a dependent search bank as descriptive unless complete executions are repeated (Hariri et al., 2026a, §3.5, App. C.2); concurrently, Singh and Singh (2026) characterise what fixed-rollout counts identify about the task-averaged independent-attempt

pass@k. We build on these. What is new here is which pipeline estimates which functional when attempts share state, the pooled functional in closed form, succ@k as the unique range-preserving estimable extension, the run-count threshold for whatever is recorded inside a run, and a valid interval below it. The degree criterion is classical (Halmos, 1946; Bickel and Lehmann, 1969; Lehmann and Casella, 1998; Kimpara et al., 2023), range-preserving unbiased estimation is Hoeffding (1984), and GDAG pairs polynomial estimability with minimax approximation for a different target (Mohri et al., 2026, §4.4, §5.1–5.4).

Replication, allocation and sequential inference. The design effect describes saturation from the variance side (Kerry and Bland, 1998; Bay and Yearick, 2026); the interval remark of Bay and Yearick (2026, §5.4) concerns the between-problem difficulty correlation that caps the effective sample size of a benchmark mean, whereas ours concerns dependence within a question and within a run. The variance floor has an inherited $k = 1$ case (Zhang et al., 2026, Lem. 2.1); allocation across questions is prior work (Kazdan et al., 2025, Thm. 1, App. D), and we price the within-question trade. We specialise Bernoulli mixture confidence sequences (Howard et al., 2021) to two run-level targets; PACE bets on discordant pairs under a conditional null (Shawn, 2026, §4); anytime-valid evaluation has precedents (Zhou et al., 2026; Ota et al., 2026) (Appendix E).

3 Three quantities

3.1 Runs, attempts, and the class of generating procedures

Fix a question. A *run* is one complete execution of the generating procedure G from its outermost source of randomness, producing attempts $X_1, \dots, X_n \in \{0, 1\}$, where $X_i = 1$ means that attempt i passes the verifier; R is the number of independent complete runs. The simplest model of shared state gives each run a hidden success rate: a good prefix makes all of its continuations likely to succeed, a bad one makes them all likely to fail.

Definition 3.1 (Conditionally i.i.d. latent-variable class)**.** For $\Pi \subseteq \mathcal{P}([0, 1])$ and $\pi \in \Pi$: draw $\theta_1, \dots, \theta_R \sim \pi$ independently; given θ_r, the attempts $X_1^{(r)}, \dots, X_n^{(r)}$ of run r are conditionally i.i.d. Bern(θ_r); runs are independent. The observation is the whole array $(X_i^{(r)})$.

The threshold of §4 holds in a wider class, $\mathcal{G}_{\mathrm{ker}}(\Pi; K, R)$, where each run is observed through an arbitrary Markov kernel K fixed in advance of π and required only to be θ-*calibrated*: the first attempt is recoverable from the observation and succeeds with probability θ_r given the run's latent draw (Definition A.1; Definition 3.1 is its $K(\theta, \cdot) = \mathrm{Bern}(\theta)^{\otimes n}$ case). Π satisfies the richness condition (T) if there exist $\theta_0 \neq \theta_1 \in [0, 1]$ and an infinite $A \subseteq [0, 1]$ with $\pi_\alpha := (1 - \alpha)\delta_{\theta_0} + \alpha\delta_{\theta_1} \in \Pi$ for every $\alpha \in A$. An *estimator* is any measurable $T : O^R \to \mathbb{R}$, O the observation space of one run, integrable throughout the class, with external randomisation permitted.

3.2 Two targets, and what pooling reports

Write $\mu_j := \mathbb{E}_\pi[\theta^j]$ and $p := \mu_1$. The independent-attempt target and the procedure-level target are

$$\mathrm{pass@}k^{\mathrm{iid}} \;:=\; 1 - (1 - p)^k, \tag{1}$$

$$\mathrm{succ@}k(G) \;:=\; \mathbb{P}(G \text{ passes within budget } k) \;=\; 1 - \mathbb{E}_\pi\big[(1 - \theta)^k\big], \tag{2}$$

the second read in the latent-variable model. The first asks how often k fresh independent attempts would succeed; the second, how often the procedure as run does. If half of the runs start from a good prefix ($\theta = 0.9$) and half from a bad one ($\theta = 0.1$), then at $k = 2$ the procedure succeeds with probability $\mathrm{succ@}k = 1 - (0.5 \cdot 0.1^2 + 0.5 \cdot 0.9^2) = 0.59$, while two independent attempts at the average rate $p = 0.5$ would succeed with probability $\mathrm{pass@}k^{\mathrm{iid}} = 0.75$: retrying inside a bad run rarely helps. Under Definition 3.1 with $k \geq 2$ the difference $\mathrm{pass@}k^{\mathrm{iid}} - \mathrm{succ@}k$ (the *estimand gap*) is non-negative and zero exactly when $\mathrm{Var}(\theta) = 0$; its second-order term is $\binom{k}{2}(1 - \mu_1)^{k-2}\,\mathrm{Var}(\theta)$ with $\mathrm{Var}(\theta) = \rho\,\mu_1(1 - \mu_1)$ for the within-run intra-class correlation ρ (Corollary A.5); outside Definition 3.1 the sign follows the direction of the dependence (Remark A.10). Given $X \in \{0, 1\}^n$ with $c = \sum_i X_i$ successes and $n \geq k$, the standard estimator of Chen et al. (2021) is

$$\widehat{C}_{n,k}(X) \;:=\; 1 - \binom{n - c}{k} \Big/ \binom{n}{k}. \tag{3}$$

A pooled pipeline applies (3) to all nR attempts of a question, giving $\widehat{C}_{\text{pooled}}$. Its expectation mixes over how a random k-subset of the pooled attempts falls across runs (Proposition A.21), and it is neither target:

Proposition 3.2 (The pooled report lies between the two estimands)**.** *Under Definition 3.1 with R runs of n attempts and $nR \geq k$,*

$$\text{succ@}k \;\leq\; \mathbb{E}[\widehat{C}_{\text{pooled}}] \;\leq\; \text{pass@}k^{\text{iid}}.$$

The first inequality is strict if and only if $\text{Var}(\theta) > 0$*, $k \geq 2$ and $R \geq 2$; the second if and only if* $\text{Var}(\theta) > 0$*, $k \geq 2$ and $n \geq 2$.*

A pooled report therefore overstates what the procedure achieves and understates $\text{pass@}k^{\text{iid}}$, strictly under the conditions above (Corollary A.23). It is not even a property of the procedure: for an unchanged procedure its expectation is $\text{succ@}k$ at $R = 1$ and tends to $\text{pass@}k^{\text{iid}}$ as $R \to \infty$ at fixed n (proof of Proposition 3.2), so adding evaluation runs changes the number reported. Outside Definition 3.1 the ordering can fail (Remark A.27).

3.3 What the classical estimator and the run indicator estimate

Proposition 3.3 (What the classical estimator is unbiased for)**.** *Let X follow an exchangeable law G on $\{0,1\}^n$ and let $n \geq k$. Then $\mathbb{E}_G[\widehat{C}_{n,k}] = \mathbb{P}_G(X_1 \vee \cdots \vee X_k = 1) = \text{succ@}k(G)$; without exchangeability it equals the subset-averaged success probability $\binom{n}{k}^{-1} \sum_{|S|=k} \mathbb{P}(\exists i \in S : X_i = 1)$.*

Applied within one run, the classical estimator thus targets the procedure, not independent attempts. Without exchangeability there is no single "single-attempt success rate" to substitute into (1): with two attempts, the first-attempt, second-attempt and pooled readings already disagree (Example F.1).

Proposition 3.4 ($\text{succ@}k$ is always estimable)**.** *For any procedure G that observes its first k attempts, $T := \mathbf{1}\{X_1 \vee \cdots \vee X_k = 1\}$ is unbiased for* $\text{succ@}k$*, with variance* $\text{succ@}k(1 - \text{succ@}k)$.

No condition is placed on the dependence or adaptivity of G: one complete run is one Bernoulli trial for procedure success. Nor is there another candidate:

Theorem 3.5 (Uniqueness of $\text{succ@}k$)**.** *Under budget-k observation $X_1, \ldots, X_k$, let F be a functional on laws of $\{0,1\}^k$ with (i) $F(G) = \mathbb{E}_G[T]$ for some $T : \{0,1\}^k \to \mathbb{R}$ and all G; (ii) $F(\text{Bern}(p)^{\otimes k}) = 1 - (1-p)^k$ for infinitely many p; and (iii) T takes values in $[0,1]$ or (iii′) T is coordinatewise non-decreasing. Then $T = \mathbf{1}\{\exists i \leq k : x_i = 1\}$ and $F = \text{succ@}k$. Dropping both (iii) and (iii′) breaks uniqueness.*

Posterior means and effective-attempt counts fall outside (i). With $n > k$ observed attempts, uniqueness gives way to the orbit-average family of Corollary A.12, all unbiased for $\text{succ@}k$ under exchangeability; a procedure that stops itself before the budget measures the truncated procedure (Propositions A.15, A.17; estimators against targets: Appendix Table 2).

4 How many runs the independent-attempt target needs

$\text{succ@}k$ needs one run. The independent-attempt target needs more, and the reason is a counting argument: each independent run contributes one draw of θ, while $\text{pass@}k^{\text{iid}}$ depends on the mixing law through a polynomial of degree k.

Theorem 4.1 (Repetition threshold; sharp)**.** *Let Π satisfy (T), let the model be $\mathcal{G}_{\text{ker}}(\Pi; K, R)$ of Definition A.1, and let $k \geq 1$. (a) An unbiased estimator of* $\text{pass@}k^{\text{iid}}$ *exists if and only if $R \geq k$. (b) For fixed R, when each run records n attempts as in Definition 3.1, no sequence of estimators T_n has $T_n \to \text{pass@}k^{\text{iid}}(\pi)$ in probability as $n \to \infty$ for all $\pi \in \Pi$.*

For $R < k$ no unbiased estimator exists however large n is and however rich the observation — complete traces, verifier scores, tree structure — and external randomisation does not help: along the two-point family the law of R runs has degree at most R in α, while the target has degree k. The criterion and the distinct-run product construction are inherited (Lehmann and Casella, 1998; Halmos, 1946; Kimpara et al., 2023); what is specific here is that the degree is counted in complete runs (Appendix F.1). At $k = 1$ the target has degree one, so pass@1 survives any dependence. The

two-point argument of part (b) also denies consistency in n at fixed R to succ@k, since more attempts per run measure only that run's draw; and averaging pass@k^{iid} over questions observed in one run each ($k \geq 2$) does not restore unbiasedness when their laws vary freely (Corollary A.13).

What the threshold costs. The threshold is a boundary of unbiased estimability, not a sample size: above it the variance still falls only as $1/R$. Below it every estimator has a worst-case bias; over classes containing the two-point family on $\{0, 1\}$ with weights dense in $[0, 1]$ the minimax bias at $R = k - 1$ is $2 \cdot 4^{-k}$, attained by an estimator of the runs' first attempts (Theorem A.6): 0.0078 at $k = 4$, against the headline design's estimand gap of 0.0262 (§6.1). Knowing that μ_1 lies in an interval of width w lowers it to $2(w/4)^k$. At $R \geq k$, with $S_r = \sum_i X_i^{(r)}$,

$$\widehat{U}_k \;:=\; 1 - \frac{(R-k)!}{R!} \sum_{(r_1,\ldots,r_k)\ \text{distinct}} \prod_{i=1}^{k} \bigl(1 - S_{r_i}/n\bigr) \tag{4}$$

is unbiased for pass@k^{iid} whenever runs are i.i.d. and every attempt has success probability p (Proposition A.8); when the first n moments of the mixing law fill an open set, as all laws on $[0, 1]$ do, no unbiased estimator has variance below $k^2(1-\mu_1)^{2(k-1)}\,\mathrm{Var}_\pi(\theta)/R$, whatever n (Propositions A.14, A.9).

5 What to report instead

Make the complete run the unit: dependence stays within runs, independence holds between them.

Protocol: a run-aware pass@k **report**, for R independent complete runs of $n \geq k$ attempts per question.

1. **Name the target**: succ@k for the procedure as run, pass@k^{iid} for the independent-attempt target; a pooled pass@k is a third quantity (Proposition 3.2).
2. succ@k**, any** R: point $\overline{Y}$, the fraction of runs that succeed within k attempts (Proposition 3.4); interval runCS of (17), valid at any stopping time, or at a fixed R a cluster-robust Wald interval (§F.6).
3. pass@k^{iid}: if $R \geq k$, $\widehat{U}_k$ of (4); if $R < k$, no uniformly unbiased point estimate exists (Theorem 4.1), and a minimax-bias estimate carries its bound (Theorem A.6); at every R, the first-attempt sequence mapped through $g_k(p) = 1 - (1-p)^k$ (§5.2).
4. **Attempts per run**: only fresh runs lower the between-run variance term V_0/R; choose $n^\star$ of Proposition A.24 (§5.4).

5.1 A run indicator absorbs all within-run dependence

For run r on question q record $Y_{q,r} := \mathbf{1}\{\text{run } r \text{ passes within budget } k\}$. The one requirement is that runs are fresh (a checklist is in Table 40):

Assumption 5.1 (Fresh runs, (I))**.** Complete run observations are independent and identically distributed. Sampled or mutable state is isolated between runs; fixed read-only inputs may be shared.

Under it $Y_{q,1}, Y_{q,2}, \ldots$ are i.i.d. Bern(s_q) with $s_q :=$ succ@$k(G, q)$, and every dependence inside a run is absorbed into Y. A pooled pipeline is repaired by applying (3) inside each run and averaging ($\widehat{C}_{\text{run}}$, unbiased for succ@k under exchangeable within-run laws) or, for ordered attempts, by the run indicator.

5.2 Run-level confidence sequences hold at any stopping time

A standard Beta-mixture e-process on run indicators gives, by Ville's inequality (Ville, 1939), the run-level sequence $\text{runCS}_t(\delta)$ of (17): it contains s_q with probability at least $1-\delta$ at any data-dependent stopping time, without within-run assumptions (Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Ramdas et al., 2023). At benchmark level, two sampling schemes reach $\bar{s} := Q^{-1}\sum_q s_q$ over Q questions without multiplicity correction, and paired comparisons reduce to the same stream on discordant pairs (Appendix F.4). The sequence is conservative: on fitted laws it covers the exact succ@k 0.997–1.000 of the time at nominal 0.95 (Tables 32, 33).

For the independent-attempt target, calibration and run independence make the first attempts $B_r := X_1^{(r)}$ i.i.d. Bern(p). Any valid interval or sequence for p maps through the strictly increasing $g_k(p) = 1-(1-p)^k$ to one for pass@k^{iid} at the same level, even without $R \geq k$; we write $g_k(\text{runCS}_1)$ for this first-attempt sequence. Across §6 it contains the held-out pass@k^{iid} reference 0.428–0.965 of the time (Table 1; containment of a held-out estimate that carries its own noise, not coverage of the target, §7); below the threshold it carries the report, at the price of width. Above the threshold, $\widehat{U}_k$ and the plug-in $g_k(\bar{X})$ have similar per-question error, but the plug-in's bias has one sign on every question, so at the benchmark level $\widehat{U}_k$ has the smaller root mean squared error in 8 of 9 design–budget cells (Table 45).

5.3 Why attempt-level intervals fail

Pooling all nR attempts into (3) fails in label and in width: it reads a third quantity as pass@k^{iid} and takes nR as the sample size. With fixed R and $\text{Var}(\theta) > 0$, $\widehat{C}_{\text{pooled}}$ converges as $n \to \infty$ to the random limit $1-(1-\bar{\theta})^k$, $\bar{\theta} := R^{-1}\sum_r \theta_r$, so any interval around it that shrinks with n has limiting coverage at most the largest atom of that limit (Proposition A.22). Question-level standard errors (Miller, 2024, §3.1) fix the standard error of a mean over questions, not the estimand under within-run dependence.

Which interval, and against what. On Arm A ($R = 32$ runs of $n = 32$ attempts on 194 questions, $k = 4$, $\delta = 0.05$), the same 16-run estimation half supports three intervals across 50 splits, scored against the other half of the runs (the held-out reference), against a fresh 128-run reference, and against the known succ@k of laws fitted to each question (Figure 2b). The attempt-level bootstrap on 512 attempts contains the fresh reference 0.104 of the time at mean width 0.052; a run-level Clopper–Pearson interval on the 16 runs contains it 0.952 of the time at width 0.386, and runCS 0.990 at width 0.502, the extra width buying validity under optional stopping. On an interval score that charges width and misses together (Winkler, 1972), run-level intervals win in 8 of 8 designs (Table 1).

Clustering widens the interval but leaves its centre. Resampling complete runs around the pooled estimate contains the held-out reference more often than the attempt-level interval in 8 of 8 designs, but where the estimand gap is material it leaves the centre off the procedure's success: on Arm A, centred on the pooled estimate, it contains the fresh reference 0.327 of the time (Table 46).

What it costs at the point of decision. Comparing two equal systems on a fixed question set, a single run's attempt-level interval certifies a nonexistent difference in 0.745 ($n = 32$) to 0.931 ($n = 512$) of simulated replicates with no run-start cost, the cluster-sampling design effect (Kerry and Bland, 1998); measured runs show the same pattern (Tables 38, 39). When the systems differ by 0.050 in question-averaged succ@4, the benchmark-level discordant-pair sequence certifies the true difference with probability 0.94–0.96 at $R = 16$, and a cluster-robust Wald interval 0.955–0.968 at $R = 8$ (both question-set panels), at type I 0.04–0.07 (Table 38; full grids in the supplementary report, Tables S22, S23 and S24).

5.4 Only fresh runs lower the variance floor

Write $g(\theta) := 1-(1-\theta)^k$, so succ@$k = \mathbb{E}_\pi[g(\theta)]$, and put

$$V_0(\pi,k) := \text{Var}_\pi\big(g(\theta)\big), \quad W_\pi(n,k) := \mathbb{E}_\pi\big[\text{Var}(\widehat{C}_{n,k}(X) \mid \theta)\big], \quad \widehat{C}_{\text{run}} := R^{-1}\textstyle\sum_r \widehat{C}_{n,k}(X^{(r)}). \tag{5}$$

Under Definition 3.1 the variance splits exactly as $\text{Var}(\widehat{C}_{\text{run}}) = [V_0(\pi,k) + W_\pi(n,k)]/R$, with V_0 free of n and $W_\pi(\cdot,k)$ non-increasing in n to zero, and, under the conditions of Proposition A.14, no unbiased estimator of succ@k has smaller variance (Theorem A.20). Attempts buy W and only fresh runs buy V_0. At fixed cost the optimum minimises $\Phi(n) := (V_0 + W_\pi(n,k))(\kappa + n)$ over $n \geq k$, where $\kappa := c_0/c_1$ is the cost c_0 of starting a run in units of the cost c_1 of one attempt, measured at 2.96 (95% interval $[0.18, 6.32]$; Table 37); on the two Beta laws of Table 5, at this κ, the relaxed optimum is 9 or 5 attempts per run, and the integer optimum at total cost $N = 128$ is 11 or 5 (Table 6).

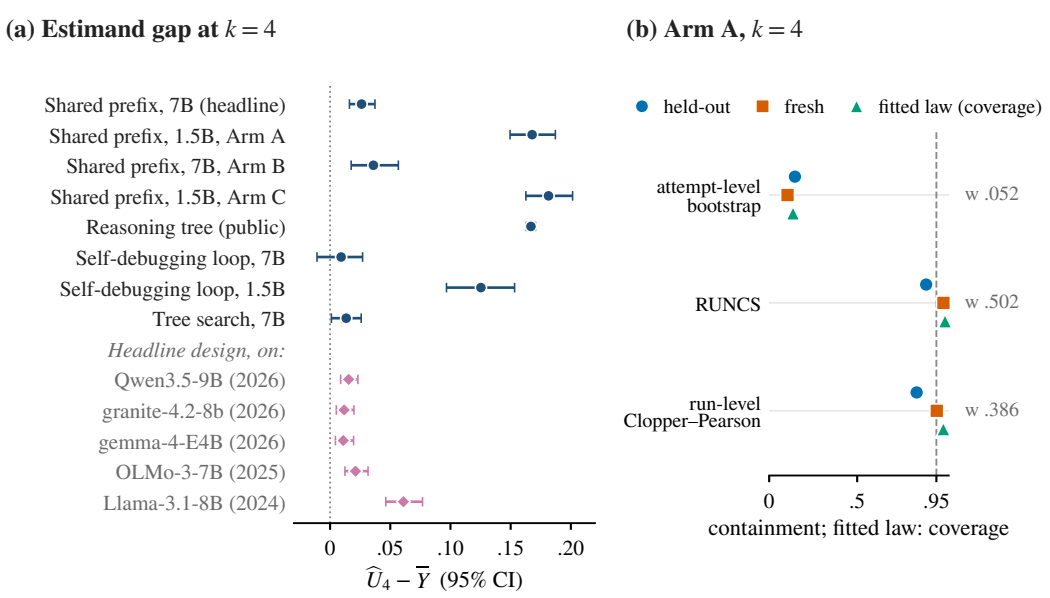


Figure 2: The estimand gap has an interval above zero in every design except the self-debugging 7B one, whose interval includes zero, and on Arm A only run-level intervals contain a fresh reference at close to the nominal rate (0.952 and 0.990, against 0.104). (a) $\widehat{U}_4 - \overline{Y}$, unbiased for pass@$k^{\text{iid}}$ − succ@4 (first-attempt form on the self-debugging and tree-search designs, where the fresh-attempt counterfactual also changes the feedback or search policy, so the gap is not dependence alone; §6.4), with 95% question-bootstrap intervals from Table 50 (the eight designs) and Table 14 (five further open-weight models). (b) Arm A, $k = 4$: containment of the held-out and fresh references and coverage of the fitted-law succ@4 by three intervals on the same estimation halves, with mean widths (Table 22).

6 Evidence

We measure the gap and the intervals on designs where we control the dependence, on a public corpus we did not generate, and on two real scaffolds. Figure 2a and Table 1 give every design at $k = 4$ (Appendix B.1 reads each row); the *headline design* is the shared-prefix 7B design of §6.1. The controlled designs use one family, Qwen2.5-Instruct (2024), at two sizes, holding tokenizer, template and sampler fixed, as do the scaffolds; five further models from four further families replicate the headline design.

6.1 A shared prefix opens a $k = 4$ estimand gap of +0.0262

Does a shared prefix alone open the predicted gap? For each of 200 GSM8K questions (Cobbe et al., 2021) we sample $R = 16$ prefixes of $P = 48$ tokens from Qwen2.5-7B-Instruct and $n = 8$ continuations of each; admissibility (no answer in the prefix, no early stop) is part of the estimand (Appendix B.2). The gap is null at $k = 1$ and +0.0262 at $k = 4$, and the finite-design expectation predicts the pooled number out of sample closely, with a small downward bias (Table 30). $\widehat{U}_4 - \overline{Y}$ is exactly zero on 158 of the 200 questions; the other 42 carry the gap, and all 200 stay in every denominator. An i.i.d. control from the same model's independent chains, run on 100 of the 200 questions, shows no gap (Table 12). Repeated unchanged on five further open-weight models (Qwen3.5-9B, granite-4.2-8b and gemma-4-E4B, 2026, thinking disabled; OLMo-3-7B, 2025; Llama-3.1-8B, 2024), the gap is +0.0154, +0.0117, +0.0110, +0.0211 and +0.0609, each with a 95% question-bootstrap interval above zero (Table 14; Appendix B.3).

6.2 A human-labelled reasoning tree carries six to seven times that gap

Does the gap persist on a corpus we did not generate? A public human-labelled reasoning tree (PRM800K; Lightman et al., 2024) supplies sibling completions of a shared prefix, with step-level labels. On the 4,880 of 8,061 questions with at least four sibling groups its gap is six to seven times

Table 1: Scored on width and misses together, run-level intervals beat attempt-level intervals in every design: the Winkler score of runCS and of the run-level Clopper–Pearson interval is below the attempt-level bootstrap's in 8 of 8 designs, and between the two run-level intervals each has the lower score in 4 and 4 designs respectively. At $k = 4$: the estimand gap $\widehat{U}_4 - \overline{Y}$ (95% interval in Figure 2a); the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ over the run-level estimate of succ@4, whose 95% interval (Table 50) lies above zero in 8 of 8 designs; containment of a held-out $\overline{Y}_4$ ($\overline{Y}$ at $k = 4$) by an attempt-level bootstrap and by runCS (17), and of the held-out statistic for pass@k^{iid} by the first-attempt sequence through g_4; mean width of the first two; and the Winkler score (Winkler, 1972) $S = (u - l) + \frac{2}{\delta}(l - y)\mathbf{1}\{y < l\} + \frac{2}{\delta}(y - u)\mathbf{1}\{y > u\}$ at $\delta = 0.05$ against the held-out $\overline{Y}_4$ (lower is better). The reasoning-tree row is on its $R_q \geq 4$ subset (Table 17). Cells are those of Tables 7, 47 and 50.

			containment			mean width		Winkler score		
Design	$\widehat{U}_4 - \overline{Y}$	$\widehat{C}_{\text{pooled}} - \overline{Y}$	attempt	runCS	g_4(runCS$_1$)	attempt	runCS	attempt	runCS	CP
Prefix, 7B (headline)	+0.0262	+0.0251	0.864	0.966	0.954	0.028	0.477	0.791	0.502	0.462
Prefix, 1.5B, Arm A	+0.1678	+0.1635	0.146	0.892	0.868	0.052	0.502	6.003	0.578	0.652
Prefix, 7B, Arm B	+0.0360	+0.0351	0.208	0.922	0.965	0.029	0.374	1.614	0.402	0.395
Prefix, 1.5B, Arm C	+0.1814	+0.1766	0.190	0.905	0.878	0.053	0.504	6.399	0.570	0.636
Reasoning tree	+0.1668	+0.1464	0.609	0.788	0.428	0.492	0.860	4.928	1.257	1.412
Self-debugging loop, 7B	+0.0092	+0.0205	0.785	0.925	0.824	0.078	0.482	0.745	0.505	0.470
Self-debugging loop, 1.5B	+0.1253	+0.0819	0.670	0.887	0.728	0.248	0.589	2.011	0.681	0.720
Tree search, 7B	+0.0134	+0.0213	0.913	0.984	0.900	0.041	0.473	0.530	0.492	0.446

the shared-prefix value; the per-step success rate, averaged over all 8,061 questions, is 0.38 rather than 0.90. Siblings are not exchangeable, so the choice of i.i.d. counterfactual matters; on matched questions and weights both unbiased readings agree in sign and size, +0.1799 from the first sibling and +0.1668 from a uniformly random one, each excluding zero (Table 18).

6.3 Away from saturation the gap stays positive up to $k = 32$

Near $\hat{\mu}_1 = 0.90$ the factor $(1 - \mu_1)^{k-2}$ compresses every dependence term, so three designs move off saturation with $R = 32$ prefixes of $n = 32$ continuations: Qwen2.5-1.5B-Instruct on 194 of the same questions (Arm A); Qwen2.5-7B-Instruct on MATH-500 questions (Hendrycks et al., 2021; Lightman et al., 2024) pre-filtered to an accuracy of 128 independent chains in $[0.05, 0.95]$ (Arm B); and Arm A without the admissibility condition (Arm C). The gap is positive at every budget $k \geq 2$; it grows to +0.2410 and +0.2596 at $k = 32$ on Arms A and C, and on Arm B peaks at $k = 8$ and falls back to +0.0287 by $k = 32$. The finite-design expectation stays tight where the second-order reference falls behind (Tables 19, 21).

6.4 Real scaffolds carry estimand gaps from null to +0.1253

Do real scaffolds, whose dependence we do not construct, carry a gap? *A self-debugging loop.* A Reflexion-style loop (Shinn et al., 2023) writes its own tests and resubmits while they fail, up to $k_{\max} = 4$ attempts, on the 164 HumanEval problems (Chen et al., 2021), with $R = 16$ fresh-seed runs on Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct (sandbox, §B.6; leakage check, Table 24(a)). Attempts are ordered, so only the first has an i.i.d. counterfactual. On the 7B model (first attempt succeeds 0.8190) the gap is indistinguishable from zero, yet the pooled report still exceeds the run-level estimate of succ@4 (Table 50); ordered attempts, here and in the tree search below, fall outside Definition 3.1, so the pooled report need not sit below pass@k^{iid} (Remark A.27). On the 1.5B model (0.5271) the estimated success probability of four fresh attempts exceeds that of the loop's four attempts by 0.1253, a contrast that also changes the feedback policy and agrees in direction with equal-cost comparisons of self-refinement against repeated sampling (Mirzaei, 2026).

A tree-search scaffold. A tree-of-thoughts search (Yao et al., 2023) on Arm A's 194 questions with Qwen2.5-7B-Instruct, with no prefix imposed and no reference answer entering generation or pruning: a run is one breadth-first search returning 4 leaves, 16 fresh runs per question. A leaf succeeds 0.8934 of the time; the gap is +0.0134 (95% interval $[+0.0011, +0.0258]$) at $k = 4$, so a success rate

near the headline design's does not by itself force a null gap; the interval still excludes zero without the runs in which a node terminated early (Tables 26, 50; Appendix B.7).

7 Limitations

Theory. The threshold concerns exact uniform unbiasedness over classes satisfying (T), per question; it is not claimed for smooth parametric families such as $\{\text{Beta}(a, b)\}$. The first-attempt sequence for pass@k^{iid} needs calibration and run independence and is read per question; g_k is strictly concave for $k \geq 2$, so a benchmark-level statement needs an interval for the averaged target. Without exchangeability pass@k^{iid} depends on which attempt is designated first — on the reasoning tree the first- and random-sibling readings differ by +0.0130 [+0.0076, +0.0186] — so an ordered design has to name that attempt; there the first-attempt sequence contains its noisier first-sibling reference only 0.428 of the time, a figure that mixes reference noise with interval error. *Validity and cost.* Run-level validity rests on Assumption 5.1; distinct seeds are operational evidence rather than proof (Table 40); the benchmark-level paired sequence assumes i.i.d. question pairs. Validity costs runs: width 0.10 at $s_q = 0.5$ takes 1,359 runs (Table 44), and at the benchmark level run-level runCS on the run indicators, unlike the discordant-pair sequence, certifies a true difference of 0.050 with probability at most 0.001 at $R \leq 8$ and 0.026 at $R = 16$, $n = 8$ ($N = 128$, $\kappa = 0$; Table 38). *Evidence.* Admissibility, Arm B's difficulty filter and the prefix cut point are part of the estimands they define. The two scaffolds are each one implementation on one benchmark with Qwen2.5-Instruct; their gap sizes describe those procedures, while the validity of the run indicator and of runCS does not depend on the model. Reasoning-tree sibling groups are assumed independent (full scope: Appendix G).

8 Conclusion

When attempts share state, the number of complete runs, not of attempts, decides what a pass@k can report. One run estimates the procedure's success; over the rich class of §3, the independent-attempt target needs k runs for a uniformly unbiased point estimate, while under calibration a run-level interval for it is valid from one; and a pooled report is a third number that moves with the evaluation design. Pooling overstated the procedure's success in all 8 designs, and on Arm A its attempt-level interval contained a fresh reference 0.104 of the time where a run-level interval contained it 0.952 of the time. For search and self-correcting evaluations the rule follows: name the target, count fresh runs, and attach a run-level interval.

AI Use Statement

We used generative AI coding and writing assistants (OpenAI Codex and Anthropic Claude) as auxiliary tools in this work: (i) to assist with implementing and debugging experimental scripts, which are not part of this article; (ii) to assist in surveying and tracking literature relevant to this study; and (iii) to improve the grammar, clarity, and readability of author-written text without altering its substantive content.

The research questions, experimental design and protocols, execution and analysis of experiments, interpretation of results, scientific claims, and final selection of references were developed and determined by the authors. All AI-assisted code was reviewed, tested, and validated by the authors before use. All literature identified with AI assistance was verified against the original sources before citation, and all AI-assisted edits to the manuscript were carefully reviewed and approved by the authors.

All reported results, analyses, and claims were independently checked by the authors for correctness, consistency, and faithfulness to the underlying experiments and evidence. Generative AI served only as an auxiliary tool and did not replace author judgment in the research process. The authors reviewed the entire manuscript and take full responsibility for its final content.

Reproducibility Statement

Proofs of every numbered result are in Appendix A; the validity of the run-level confidence sequence follows from Ville's inequality (§F.4). The protocol, corpora and freshness checklist are in Appendix D, and §D.4 lists the supplementary material: `passk_core.py` implements the estimators, intervals and the exact finite expansion; `repro/` holds the ordered attempt outcomes, split assignments and seeds behind Tables 22 and 23 and the per-question readings behind the numeric rows of Table 13, with a script that rebuilds those cells exactly and one that recomputes Table 4 by exact enumeration; and `run-provenance.md` maps each run label to the tables that read it. The package replays every analysis from the recorded pass/fail outcomes; it does not regenerate the model outputs or re-run the verifiers.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396, 2023. doi: 10.18653/v1/2023.emnlp-main.761.

Yong Yi Bay and Kathleen A. Yearick. When More Sampling Hurts: The Modal Ceiling and Correlation Ceiling of Test-Time Scaling, 2026. URL https://arxiv.org/abs/2606.28661.

P. J. Bickel and E. L. Lehmann. Unbiased Estimation in Convex Families. *The Annals of Mathematical Statistics*, 40(5):1523–1535, 1969. doi: 10.1214/aoms/1177697370.

Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On Randomness in Agentic Evals. ICLR 2026 Workshop on Agents in the Wild, 2026. URL https://arxiv.org/abs/2602.07150v3.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, 2024. URL https://arxiv.org/abs/2407.21787.

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, 2021. URL https://arxiv.org/abs/2107.03374.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching Large Language Models to Self-Debug. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KuPixIqPiq.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168.

Paula Cordero-Encinar and Andrew B. Duncan. Certified Self-Consistency: Statistical Guarantees and Test-Time Training for Reliable Reasoning in LLMs, 2025. URL https://arxiv.org/abs/2510.17472.

Pierluca D'Oro, Sneha Silwal, William Wong, Yuxuan Sun, Fanyi Xiao, Manchen Wang, Eric Gan, Allen Bolourchi, and Joseph Tighe. Computer Use at the Edge of the Statistical Precipice, 2026. URL https://arxiv.org/abs/2605.08261.

Bodong Du, Xuanqi Huang, and Xiaomeng Li. Distribution-Aware Reward Estimation for Test-Time Reinforcement Learning, 2026. URL https://arxiv.org/abs/2601.21804.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, July 2014. doi: 10.1002/qj.2270.

Paul R. Halmos. The Theory of Unbiased Estimation. *The Annals of Mathematical Statistics*, 17(1): 34–43, March 1946. ISSN 0003-4851. doi: 10.1214/aoms/1177731020. URL http://dx.doi.org/10.1214/aoms/1177731020.

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, 2026a. URL https://arxiv.org/abs/2608.04001.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't Pass@k: A Bayesian Framework for Large Language Model Evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=PTXi3Ef4sT.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, September 1948. ISSN 0003-4851. doi: 10.1214/aoms/1177730196. URL http://dx.doi.org/10.1214/aoms/1177730196.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Wassily Hoeffding. Range Preserving Unbiased Estimators in the Multinomial Case. *Journal of the American Statistical Association*, 79(387):712–714, 1984. doi: 10.1080/01621459.1984.10478101.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, April 2021. ISSN 0090-5364. doi: 10.1214/20-aos1991. URL http://dx.doi.org/10.1214/20-aos1991.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation, 2026. URL https://arxiv.org/abs/2601.20802.

Md. Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. MapCoder: Multi-Agent Code Generation for Competitive Problem Solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4912–4944, 2024. doi: 10.18653/v1/2024.acl-long.269.

Jiajun Jiang, Sharon Zheng, Natan Vidra, and Spurthi Setty. Beyond Pass@k: Measuring Reliability and Security of Agentic Code Generation, 2026. URL https://arxiv.org/abs/2608.14711.

Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient Prediction of Pass@k Scaling in Large Language Models, 2025. URL https://arxiv.org/abs/2510.05197.

Sally M. Kerry and J. Martin Bland. The intracluster correlation coefficient in cluster randomisation. *BMJ*, 316(7142):1455, 1998. doi: 10.1136/bmj.316.7142.1455.

Dhamma Kimpara, Rafael Frongillo, and Bo Waggoner. Proper Losses for Discrete Generative Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pages 17015–17040, 2023.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, 2nd edition, 1998. ISBN 0387985026. doi: 10.1007/b98854. URL https://doi.org/10.1007/b98854.

Noam Levi. A Simple Model of Inference Scaling Laws, 2024. URL https://arxiv.org/abs/2410.16377.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. QuasiMoTTo: Quasi-Monte Carlo Test-Time Scaling, 2026. URL https://arxiv.org/abs/2607.01179.

Xiao Liang, Zhong-Zhi Li, Zhenghao Lin, Eric Hancheng Jiang, Hengyuan Zhang, Yelong Shen, Kai-Wei Chang, Ying Nian Wu, Yeyun Gong, and Weizhu Chen. Training LLMs for Divide-and-Conquer Reasoning Elevates Test-Time Scalability, 2026. URL https://arxiv.org/abs/2602.02477.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Licheng Liu, Zihan Wang, Linjie Li, Chenwei Xu, Yiping Lu, Han Liu, Avirup Sil, and Manling Li. A Simple "Try Again" Can Elicit Multi-Turn LLM Reasoning, 2025. URL https://arxiv.org/abs/2507.14295.

Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Iliya Mirzaei. Sample More, Reflect Less: Self-Refine and Reflexion Lose to Repeated Sampling at Equal Token Cost, from 1.5B to 7B, 2026. URL https://arxiv.org/abs/2607.28576.

Mehryar Mohri, Jon Schneider, and Yutao Zhong. Generalized Distributional Alignment Games for Unbiased Answer-Level Fine-Tuning, 2026. URL https://arxiv.org/abs/2605.02435.

Zairah Mustahsan, Abel Lim, Megna Anand, Saahil Jain, and Bryan McCann. Stochasticity in Agentic Evaluations: Quantifying Inconsistency with Intraclass Correlation, 2025. URL https://arxiv.org/abs/2512.06710.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. CITE: Anytime-Valid Statistical Inference in LLM Self-Consistency, 2026. URL https://arxiv.org/abs/2605.05873.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025. URL https://arxiv.org/abs/2412.15115.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4), 2023. doi: 10.1214/23-STS894.

Tal Ridnik, Dedy Kredo, and Itamar Friedman. Code Generation with AlphaCodium: From Prompt Engineering to Flow Engineering, 2024. URL https://arxiv.org/abs/2401.08500.

Jongha Jon Ryu and Gregory W. Wornell. Gambling-Based Confidence Sequences for Bounded Random Vectors. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 42856–42869, 2024.

Zayx Shawn. PACE: Anytime-Valid Acceptance Tests for Self-Evolving Agents, 2026. URL https://arxiv.org/abs/2606.08106.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pranav Singh and Prashant Singh. What Fixed-Rollout pass@k Evaluations Can Identify, 2026. URL https://arxiv.org/abs/2609.09245.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Phat T. Tran-Truong and Xuan-Bach Le. Measuring the Unmeasurable: Markov Chain Reliability for LLM Agents, 2026. URL https://arxiv.org/abs/2604.24579.

Jean Ville. Étude critique de la notion de collectif. Thèse de doctorat ès sciences mathématiques, Faculté des Sciences de Paris (soutenue le 9 mars 1939); Monographies des Probabilités, Calcul des Probabilités et ses Applications, dir. Émile Borel, Fascicule III, Gauthier-Villars, Paris, 1939. URL https://www.numdam.org/item/THESE_1939__218__1_0.pdf.

Christian Walder and Deep Karkhanis. Pass@K Policy Optimization: Solving Harder Reinforcement Learning Problems, 2025. URL https://arxiv.org/abs/2505.15201.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Robert L. Winkler. A Decision-Theoretic Approach to Interval Estimation. *Journal of the American Statistical Association*, 67(337):187–191, 1972. doi: 10.1080/01621459.1972.10481224.

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, 2023.

Wenbo Zhang, Hengrui Cai, and Wenyu Chen. Beyond the Singular: Revealing the Value of Multiple Generations in Benchmark Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 10033–10043, 2026. doi: 10.18653/v1/2026.findings-acl.488.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 62138–62160, 2024.

Zhijian Zhou, Zesheng Ye, Zhaorun Chen, Bo Li, and Feng Liu. CELEUS: Certifiable and Efficient LLM Evaluation via E-Processes, 2026. URL https://arxiv.org/abs/2606.20820.

Guide to the appendix

The table lists each appendix, the question it answers and the main-text passage it supports. The supplementary report (`supplement.pdf` in the supplementary material; sections and tables numbered S1, S2, …) holds the full grids, robustness checks and material no main-text sentence rests on.

Appendix	Question it answers	Supports
A Proofs	What exactly is claimed, and why is it true?	every result of §§3–5
B Evidence	Where does each number of the evidence come from, mechanism by mechanism?	§6, Table 1, Figure 2
C Simulation and calibration	Do the predictions hold against a known truth; how much do noisy references move containment; what happens at a decision?	§5, including the decision paragraph of §5.3
D Data and protocol	How were the runs made and checked, what does practice compute, and what does the supplementary material contain?	the practice audit of §1; the freshness checklist; the Reproducibility Statement
E Related work	How does each neighbouring result differ?	§2
F Extended arguments	What are the longer forms of arguments stated briefly?	the pointers of §§3–5
G Extended scope	What is not claimed?	§7

A Proofs

This appendix states every result the main text uses in full (§A.1) and proves each one, section by section (§§A.2–A.4).

A.1 Full statements of the main-text results

The statements below are used in the body and proved later in this appendix.

From Section 3.

Definition A.1 (General latent-variable class)**.** Fix a Markov kernel $K : [0,1] \to O$ into an observation space O, not depending on π. Draw $\theta_1, \dots, \theta_R \sim \pi$ i.i.d., let $O_r \mid \theta_r \sim K(\theta_r, \cdot)$ independently across runs, and observe $(O_1, \dots, O_R)$. We require K to be *θ-calibrated*: $X_1^{(r)}$ is a measurable function of O_r and $\mathbb{P}(X_1^{(r)} = 1 \mid \theta_r) = \theta_r$. We write this class $\mathcal{G}_{\mathrm{ker}}(\Pi; K, R)$; Definition 3.1 is the special case $K(\theta, \cdot) = \mathrm{Bern}(\theta)^{\otimes n}$.

Assumption A.2 (Nested budgets, (N))**.** The procedure family $\{G_k\}$ satisfies: the attempt sequence of G_k is the first k attempts of G_n for every $k \le n$.

Assumption A.2 makes "succ@k for several k from one run" well defined; a self-correcting agent whose transitions read the budget violates it and is covered by Proposition 3.4 at each budget separately, but not by the allocation results of §5.4.

Lemma A.3 (Characterisation of estimable functionals from one run)**.** *Let $R = 1$ and let the model be Definition 3.1, with any class Π of mixing laws. For any $T : \{0,1\}^n \to \mathbb{R}$ we have $\mathbb{E}_\pi[T] = \mathbb{E}_\pi[B_c(\theta)]$ with $B_c(\theta) = \sum_{s=0}^n c_s \binom{n}{s} \theta^s (1-\theta)^{n-s}$, where c_s is the average of T on $\{|x| = s\}$; the map $c \mapsto B_c$ is a linear bijection from $\mathbb{R}^{n+1}$ onto polynomials of degree at most n. Consequently the set of functionals unbiasedly estimable from one run is exactly $\{\pi \mapsto \sum_{j=0}^n a_j \mu_j\}$.*

Corollary A.4 (Optimality at the correct target)**.** *Under Definition 3.1 with $R = 1$, $\Pi \supseteq \{\delta_\theta : \theta \in \Theta\}$ for infinite Θ, and $n \ge k$: the success count $S = \sum_i X_i$ is complete sufficient and $\widehat{C}_{n,k}$ is the UMVUE of* succ@k.

Corollary A.5 (Exact form of the gap)**.** *Under Definition 3.1 with $k \ge 2$, writing $\bar{u} = 1 - \mu_1$ and $\gamma_j = \mathbb{E}[(\theta - \mu_1)^j]$,*

$$\mathrm{pass@}k^{\mathrm{iid}} - \mathrm{succ@}k \;=\; \sum_{j=2}^{k} \binom{k}{j} \bar{u}^{\,k-j} (-1)^j \gamma_j, \qquad \mathrm{succ@}k - \mathrm{pass@}k^{\mathrm{iid}} \le 0 \text{ with equality iff } \mathrm{Var}(\theta) = 0, \tag{6}$$

with second-order approximation $\binom{k}{2}(1-\mu_1)^{k-2}\,\mathrm{Var}(\theta)$ and $\mathrm{Var}(\theta) = \rho\,\mu_1(1-\mu_1)$, where $\rho = \mathrm{corr}(X_1, X_2)$ is the within-run intra-class correlation.

Table 2: Each estimator is unbiased for one of the quantities it gets read as (for $\widehat{C}_{\text{run}}$, succ@k and the subset average, which coincide under exchangeability). Rows are the targets, columns the estimators; "low" and "high" are signed by (6) inside Definition 3.1.

Target	$\widehat{C}_{\text{run}}$	$\widehat{C}_{\text{pooled}}$	$\widehat{U}_k$
pass@k^{iid}, (1)	low, by (6)	low (Cor. A.23)	unbiased (defined only at $R \geq k$)
succ@k, (2)	unbiased (Prop. 3.3)	high when $\operatorname{Var}\theta > 0$, $k \geq 2$ and $R > 1$	high, by (6)
$\mathbb{E}[\widehat{C}_{\text{pooled}}]$, (10)	—	unbiased by construction	—
subset average, Prop. 3.3	unbiased without exchangeability	—	—

Table 3: From a single run ($R = 1$) no estimator of pass@k^{iid}, of any form and from any observation, has a smaller worst-case absolute bias than these values, which grow with the budget and tend to $1/2$; a function of the runs' first attempts attains them. Minimax bias of Theorem A.6(c) over any class of mixing laws that contains the two-point family on $\{0, 1\}$ with weights dense in $[0, 1]$ — for instance when nothing restricts the mixing law. Knowledge that confines μ_1, or a known bound on $\operatorname{Var}(\theta)$, can lower it. The row is the closed form $\frac{1}{2}(k^{-1/(k-1)} - k^{-k/(k-1)})$. Analytical values, verified in exact rational arithmetic; no experimental input.

Budget k	2	3	4	6	8	12	16	32
Minimax worst-case |bias|	0.125	0.192	0.236	0.291	0.325	0.366	0.390	0.433

The remainder beyond the second-order term has an explicit bound, recorded with the proof; in the measured regimes it is not informative, so the second-order expression serves as a magnitude reference and Proposition A.21 as the finite-design prediction.

Interpreting the moment expansion. The exact expansion is (6). Plainly: the leading term of the gap is proportional to $\operatorname{Var}(\theta)$, the variance of run difficulty; the gap is second order in how much run difficulty varies and zero exactly when it does not.

From Section 4.

Theorem A.6 (Minimax bias below the threshold)**.** *Let $R < k$, write $\psi(q) := 1 - (1 - q)^k$, and for an interval $I = [a, b] \subseteq [0, 1]$ with $a < b$ let $d_R(I) := \min_{p \in \mathcal{P}_R} \max_{q \in I} |p(q) - \psi(q)|$, the uniform distance on I from ψ to the polynomials $\mathcal{P}_R$ of degree at most R. Work in $\mathcal{G}_{\text{ker}}(\Pi; K, R)$ of Definition A.1. (a)* Lower bound. *If Π contains $\pi_\alpha = (1 - \alpha)\delta_a + \alpha\delta_b$ for every α in a dense subset of $[0, 1]$, every estimator T has $\sup_{\pi \in \Pi} |\mathbb{E}_\pi[T] - \text{pass@}k^{\text{iid}}(\pi)| \geq d_R(I)$. (b)* Attainment through μ_1 alone. *There is an estimator $T^\star$, a function of the runs' first attempts, whose bias $\mathbb{E}_\pi[T^\star] - \text{pass@}k^{\text{iid}}(\pi) = p^\star(\mu_1) - \psi(\mu_1)$ depends on π only through μ_1, where $p^\star$ is the best uniform approximant to ψ on I in $\mathcal{P}_R$; its absolute bias is at most $d_R(I)$ at every law of the class with $\mu_1 \in I$. Hence for every Π that contains the family of (a) and whose laws all have $\mu_1 \in I$, the minimax bias over Π equals $d_R(I)$. (c)* Values. $d_{k-1}(I) = 2\big((b-a)/4\big)^k$. *With $I = [0, 1]$ — nothing known about μ_1, as over all mixing laws — this is $2 \cdot 4^{-k}$ (0.0078 at $k = 4$, 0.125 at $k = 2$), and $d_1([0, 1]) = \frac{1}{2}\big(k^{-1/(k-1)} - k^{-k/(k-1)}\big)$ (Table 3).*

The equality in (b) holds over classes that contain the two-point family on the endpoints of I; with $I = [0, 1]$ these are the classes containing the two-point family on $\{0, 1\}$, all mixing laws among them. A known bound $\operatorname{Var}(\theta) \leq v$ is a different restriction, which excludes that family, and under it estimators that read several attempts of a run can do better than any function of first attempts: at $k = 2$ and $R = 1$, pass@2^{iid} = succ@2 + $\operatorname{Var}(\theta)$, so under Definition 3.1 with $n \geq 2$ the within-run estimator (3) of succ@2 plus $v/2$ has bias at most $v/2$ in absolute value. The density in (a) is needed and is used only there; Theorem 4.1 uses only that two polynomials agreeing on an infinite set agree everywhere.

Remark A.7 (One estimator family on both sides of the threshold, and what crossing it buys at equal cost). With $W_r := S_r/n$ (or, under a calibrated kernel, the run's first attempt) and e_i the ith elementary symmetric polynomial, independence across runs gives $\mathbb{E}_\pi[e_i(W)/\binom{R}{i}] = \mu_1^i$ for every law π of

Definition 3.1 and every $i \le R$, so for any polynomial $p(q) = \sum_{i \le R} a_i q^i$,

$$L[p] \;:=\; \sum_{i=0}^{R} a_i \,\frac{e_i(W)}{\binom{R}{i}} \qquad \text{has} \qquad \mathbb{E}_\pi\big[L[p]\big] \;=\; p(\mu_1) \quad \text{exactly.} \tag{7}$$

Expanding $\sum_{|J|=k} \prod_{r\in J}(1-W_r) = \sum_j (-1)^j e_j(W)\binom{R-j}{k-j}$ with $\binom{R-j}{k-j}\big/\binom{R}{k} = \binom{k}{j}\big/\binom{R}{j}$ gives $\widehat{U}_k = L[\psi]$ term for term whenever $R \ge k$, so the threshold is exactly the point at which ψ, of degree k, stops being admissible as p; at $R = k-1$ the best admissible p on I is the Chebyshev correction $p^\star = \psi - (-1)^{k+1}\,2\big(|I|/4\big)^k\,T_k\big((q-c)/h\big)$ with c the midpoint and h the half-width of I. Table 4 sets $L[p^\star]$ at $R = k-1$ against $\widehat{U}_k$ at $R = k$ under a common cost $R(c_0 + nc_1)$, with both moments computed by exact enumeration. With μ_1 unrestricted and on Arm A's interval the unbiased design has the smaller worst-case root mean squared error, by up to 1.195×; on an interval of width 0.287 around the headline design's mean, a restriction a measured dispersion does not supply, the low-bias design comes in at 0.497–0.931 times the unbiased design's risk (18 of 54 cells, all on that interval), with worst-case bias 5.29×10^{-5} at $k = 4$.

Proposition A.8 (Constructive estimator for $R \ge k$)**.** *Let $R \ge k$ runs of n attempts be i.i.d., with every attempt succeeding with the same probability p (Definition 3.1 is such a case, with $p = \mu_1$; no condition is placed on the dependence inside a run). The statistic $\widehat{U}_k$ of* (4)*, formed from the run success counts $S_r = \sum_i X_i^{(r)}$, is unbiased for $1 - (1-p)^k$.*

Proposition A.9 (Variance floor: the effective sample size is R)**.** *In the setting of Proposition A.14 with $R \ge k$, every unbiased estimator T of* pass@k^{iid} *satisfies, for every π,*

$$\mathrm{Var}_\pi(T) \;\ge\; \mathrm{Var}_\pi(\widehat{U}_k) \;\ge\; \mathrm{Var}_\pi(U_k^\theta) \;\ge\; \frac{k^2}{R}\,(1-\mu_1)^{2(k-1)}\;\mathrm{Var}_\pi(\theta), \tag{8}$$

where $U_k^\theta := 1 - \frac{(R-k)!}{R!}\sum_{(r_i)\ \text{distinct}} \prod_{i\le k}(1-\theta_{r_i})$. The right-hand side does not depend on n.

Remark A.10 (Sign discipline). Equation (6) holds inside the conditionally-i.i.d. class of Definition 3.1. Outside it the inequality can reverse, but the condition under which it does is stronger than a negative pairwise correlation. For attempts with a common marginal $\mathbb{P}(X_i = 1) = \mu_1$ under *negative lower-orthant dependence* — for every set S of attempts, $\mathbb{P}(X_i = 0\ \forall i \in S) \le \prod_{i\in S}\mathbb{P}(X_i = 0)$ — the probability that all k attempts fail is at most $(1-\mu_1)^k$, so succ@$k \ge$ pass@k^{iid}; at $k = 2$ that condition is exactly $\mathrm{Cov}(X_1, X_2) \le 0$, so a non-positive pairwise covariance already suffices there. It does not suffice for $k \ge 3$: the exchangeable law on three attempts whose success count takes the values $0, 1, 2, 3$ with probabilities $0.2, 0.2, 0.6, 0$ has $\mu_1 = 7/15$ and pairwise correlation $-1/14$, yet succ@$3 = 0.800$ against $1-(1-\mu_1)^3 = 0.848$. The sign of the gap is therefore set by the direction of the dependence: positive dependence such as plan locking or mode collapse makes succ@k smaller than pass@k^{iid}, negative lower-orthant dependence makes it larger.

Corollary A.11 (Single-run reporting)**.** *Let Π satisfy (T). With $R = 1$, no unbiased estimator of* pass@k^{iid} *exists for any $k \ge 2$, regardless of n.*

Corollary A.12 (Rigidity)**.** *Let $R = 1$, $n \ge k$ and $\Pi \supseteq \{\delta_\theta : \theta \in \Theta\}$ with Θ infinite. If $T : \{0,1\}^n \to \mathbb{R}$ is unbiased for $1-(1-\theta)^k$ on the i.i.d. submodel $\{\mathrm{Bern}(\theta)^{\otimes n}\}$, then (a) its orbit average is $c_s = 1 - \binom{n-s}{k}\big/\binom{n}{k}$, that is, exactly* (3)*; and (b) for* every *exchangeable law G, $\mathbb{E}_G[T] =$ succ@$k(G)$.*

Corollary A.13 (Benchmark averaging does not help)**.** *Suppose a benchmark reports $Q^{-1}\sum_q$ pass@$k^{\mathrm{iid}}(\pi_q)$ with $R_q = 1$ run per question, independent questions, and question laws that vary independently — the class of benchmark laws is a product $\Pi_1 \times \cdots \times \Pi_Q$. If at least one Π_{q^*} satisfies (T), then for $k \ge 2$ no estimator is unbiased for the average over the product class.*

Proposition A.14 (Complete sufficiency and optimality)**.** *Under Definition 3.1 with R runs of n attempts, suppose the moment image $\{(\mu_1, \ldots, \mu_n)(\pi) : \pi \in \Pi\}$ has non-empty interior in $\mathbb{R}^n$. Let $\mathbf{N} = (\mathbf{N}_0, \ldots, \mathbf{N}_n)$ with $\mathbf{N}_s := \#\{r : S_r = s\}$. Then $\mathbf{N}$ is complete sufficient, and hence (i) every estimable functional has at most one $\mathbf{N}$-measurable unbiased estimator; (ii) for $R \ge k$, $\widehat{U}_k$ is the UMVUE of* pass@k^{iid}*; (iii) for $n \ge k$, $R^{-1}\sum_r \widehat{C}_{n,k}(X^{(r)})$ is the UMVUE of* succ@k*.*

Write τ for the time at which a run stops, $H_\tau := (X_1, \ldots, X_\tau)$ for the *stopped history* — the attempts the run actually made, whose length is τ — and, for the first-success rule $\tau := \min\{i \le N_{\max} : X_i = 1\}$ with $\tau := N_{\max}$ when no attempt succeeds, $W := (\tau, X_\tau)$ for the success-truncated observation. Under that rule every attempt before τ is a failure, so W and H_τ determine one another; under a general rule they do not, and the stopped history is what an execution leaves behind.

Proposition A.15 (Success-truncated sampling loses nothing, and fixes everything)**.** *Let $R = 1$ and let $\Pi \supseteq \{\delta_\theta : \theta \in \Theta\}$ for an infinite Θ — the richness Corollary A.4 asks for, satisfied in particular by the class of all mixing laws and by any class containing infinitely many point masses. Then the estimable functionals from W are exactly those estimable from the full attempt vector, and each has a* unique *unbiased estimator; that of μ_1 is $\mathbf{1}\{\tau = 1,\ X_\tau = 1\} = X_1$, and that of* succ@$k$ *(for $N_{\max} \geq k$) is $\mathbf{1}\{\tau \leq k,\ X_\tau = 1\}$, which is unbiased over the whole latent-variable class.*

Remark A.16 (Where the richness hypothesis is used, and what fails without it). Unbiasedness of the two named estimators needs no hypothesis on Π: both are solved at the level of the polynomial h in the proof, so they are unbiased for every mixing law. Only *uniqueness* uses the hypothesis, and it fails without it. Take $N_{\max} = 2$ and $\Pi = \{(1-\alpha)\delta_{1/4} + \alpha\delta_{3/4} : \alpha \in (0,1)\}$, a class that satisfies richness (T) — the weight set $(0,1)$ is infinite — and contains no point mass, the endpoints being excluded. The stopped observation takes the three values $(1,1), (2,1), (2,0)$; setting $h := 3, -13, 3$ on them gives $\mathbb{E}_\pi[h] = 0$ for every $\pi \in \Pi$, since the law of W is affine in α and the two endpoints give $\frac{3}{4} - \frac{39}{16} + \frac{27}{16} = 0$ at $\theta = 1/4$ and $\frac{9}{4} - \frac{39}{16} + \frac{3}{16} = 0$ at $\theta = 3/4$. Then T and $T + h$ are two different unbiased estimators of the same functional. A two-point family leaves the law of one run only two affine degrees of freedom, which is the mechanism of Theorem 4.1 seen from the other side; uniqueness needs the $N_{\max} + 1$ independent evaluations that infinitely many point masses supply.

Proposition A.17 (Deterministic stopping rules that break identification)**.** *Take the general procedure class of Proposition 3.4: arbitrary dependence among $X_1, \ldots, X_{N_{\max}}$ and, in particular, no restriction on the conditional law of the attempts a rule leaves unrun. Let τ be* deterministic*, that is, a function of the observed history alone with no external randomisation, and suppose it stops before attempt k without a success with positive probability, $\mathbb{P}(\tau < k,\ X_1 = \cdots = X_\tau = 0) > 0$. Then* succ@$k$ *of the potential sequence is not identified from the stopped history H_τ: two laws in the class induce the same law of H_τ and different values of* succ@k*. What H_τ does identify is the success probability of the procedure as truncated, $\mathbb{P}(\exists\, i \leq \tau \wedge k : X_i = 1)$, which is* succ@$k$ *of the truncated procedure and is estimable by Proposition 3.4; the two quantities coincide whenever the rule never stops early without a success. The history is the right observation to state this in: outside the first-success rule of Proposition A.15 the pair (τ, X_τ) need not identify even the truncated success probability (Remark A.18(iii)).*

Remark A.18 (Randomised rules, and the exchangeable class). Both restrictions are used, and the proposition is a statement about deterministic rules on the general class rather than about early stopping as such. (i) *Randomisation can restore identification.* Take $k = 2$ and the rule that runs X_1 always and, after $X_1 = 0$, tosses a fair coin to decide whether to run X_2. Every history is then continued with known positive probability, so inverse-probability weighting applies: $\mathbf{1}\{X_1 = 1\} + 2\,\mathbf{1}\{X_1 = 0,\ \tau = 2,\ X_2 = 1\}$ is unbiased for succ@2 under any joint law, and the same Horvitz–Thompson construction covers any rule whose continuation probability is known and positive at every history. (ii) *Inside the exchangeable class a deterministic rule can also be enough.* Under Definition 3.1 with $k = 2$, the rule that stops at the first failure and continues after a success — which does stop early without a success — produces the terminal histories (0), $(1,0)$ and $(1,1)$ with probabilities $\mathbb{E}[1-\theta]$, $\mathbb{E}[\theta(1-\theta)]$ and $\mathbb{E}[\theta^2]$. Those three probabilities determine $\mathbb{E}[\theta]$ and $\mathbb{E}[\theta^2]$, and succ@2 $= 2\mathbb{E}[\theta] - \mathbb{E}[\theta^2]$ lies in their linear span, so it is identified from H_τ and estimated without bias by $2\,\mathbf{1}\{H_\tau = (1,0)\} + \mathbf{1}\{H_\tau = (1,1)\}$. The proposition's force comes from the general class, where the conditional law of the unrun attempts is unconstrained and no such moment structure is available. (iii) *The history does not compress to (τ, X_τ).* Take $N_{\max} = k = 3$ and the constant rule $\tau \equiv 2$, which is deterministic and stops before attempt k without a success with positive probability. Let $X_2 = X_3 = 0$ almost surely under both of two laws in the general class: under P, $X_1 \sim \mathrm{Bern}(1/2)$; under P', $X_1 = 0$ almost surely. Then $(\tau, X_\tau) = (2,0)$ with probability one under each, while the truncated success probability $\mathbb{P}(X_1 \vee X_2 = 1)$ is $1/2$ under P and 0 under P'. The stopped histories do separate them — $(1,0)$ and $(0,0)$ with probability $1/2$ each against $(0,0)$ with probability one — and the truncated success indicator is a function of H_τ by construction. This is why Proposition A.17 is stated in H_τ and Proposition A.15, where the two observations agree, in W.

Remark A.19 (Degree, and what degree does not reach). Nothing uses the target beyond its degree. Along π_α the law of R runs is a product of R mixtures each affine in α, so $\alpha \mapsto \mathbb{E}_{\pi_\alpha}[T]$ is a polynomial of degree at most R for every estimator T: a functional whose restriction to π_α is a polynomial of degree d in α admits an unbiased estimator from $\mathcal{G}_{\ker}(\Pi; K, R)$ only if $R \geq d$, whatever n and K, and a functional whose restriction is not a polynomial at all admits none at any fixed finite R. The latent variance $\mathrm{Var}_\pi(\theta)$ has $d = 2$, so it needs two runs and no number of attempts inside one will do instead. The argument does not transfer to the intra-class correlation $\rho_\pi = \mathrm{Var}_\pi(\theta)/(\mu_1(1-\mu_1))$,

which is a ratio of such functionals rather than one of them: on $\pi_\alpha = (1-\alpha)\delta_{1/4} + \alpha\delta_{3/4}$ it equals $(4\alpha - 4\alpha^2)/(3+4\alpha-4\alpha^2)$, and no polynomial Q satisfies $Q(\alpha)(3+4\alpha-4\alpha^2) = 4\alpha - 4\alpha^2$ — degrees force Q constant, and then the constant and linear coefficients demand $3Q = 0$ and $4Q = 4$. Over a class containing that family, therefore, no fixed finite number of runs admits a uniformly unbiased estimator of ρ_π; two runs buy the variance, not the correlation, and the $\hat{\rho}$ of Appendix B is a plug-in ratio reported as such.

From Section 5.

Theorem A.20 (Variance allocation)**.** *Let a run consist of a top-level randomisation Z and $n \geq k$ continuations $X_1, \dots, X_n$ drawn under it, let the complete run vectors $(Z_r, X^{(r)})$, $r = 1, \dots, R$, be i.i.d. (Assumption 5.1 with the run so defined), and let the continuations be exchangeable given Z. Then $\mathbb{E}[\widehat{C}_{n,k} \mid Z] = \mathbb{P}(X_1 \vee \cdots \vee X_k = 1 \mid Z)$, $\widehat{C}_{\text{run}}$ is unbiased for* succ@k*, and*

$$\operatorname{Var}(\widehat{C}_{\text{run}}) \;=\; \frac{V_0 + W}{R}, \qquad V_0 := \operatorname{Var}\big(\mathbb{P}(X_1 \vee \cdots \vee X_k = 1 \mid Z)\big), \quad W := \mathbb{E}\big[\operatorname{Var}(\widehat{C}_{n,k} \mid Z)\big]. \tag{9}$$

Under Definition 3.1 (the case $Z = \theta$) these are $V_0(\pi, k)$ and $W_\pi(n, k)$ of (5)*: V_0 does not depend on n, and $W_\pi(\cdot, k)$ is non-increasing in n with $W_\pi(n, k) \to 0$; and under the conditions of Proposition A.14, every unbiased estimator T of* succ@k *satisfies $\operatorname{Var}_\pi(T) \geq \operatorname{Var}_\pi(\widehat{C}_{\text{run}}) \geq V_0(\pi, k)/R$ for every $\pi \in \Pi$.*

Proposition A.21 (Expectation of the pooled classical estimator over R runs)**.** *Under Definition 3.1 with R runs of n attempts each, index the pooled attempts by pairs (r, i), write $M = nR \geq k$ and let D be the pooled number of successes, so that a pooled pipeline reports $\widehat{C}_{\text{pooled}} := 1 - \binom{M-D}{k} \big/ \binom{M}{k}$. Put $m_c := \mathbb{E}_\pi\big[(1-\theta)^c\big]$, so that $m_0 = 1$ and $m_1 = 1 - \mu_1$. Then*

$$\mathbb{E}[\widehat{C}_{\text{pooled}}] \;=\; 1 - \mathbb{E}_c\Big[\prod_{r=1}^{R} m_{c_r}\Big], \qquad \mathbb{P}(c) \;=\; \binom{M}{k}^{-1} \prod_{r=1}^{R} \binom{n}{c_r}, \tag{10}$$

where $c = (c_1, \dots, c_R)$ with $\sum_r c_r = k$ is the multivariate hypergeometric allocation across runs of a uniformly random k-subset of the pooled attempts.

Proposition A.22 (Coverage collapse)**.** *Under Definition 3.1 with R fixed and $\operatorname{Var}(\theta) > 0$, as $n \to \infty$: (a) $\widehat{C}_{\text{pooled}}$ converges in probability to a* random *limit L with $\operatorname{Var}(L) > 0$, namely $L = 1 - (1-\bar{\theta})^k$ with $\bar{\theta} := R^{-1}\sum_r \theta_r$; (b) for any interval I_n that contains $\widehat{C}_{\text{pooled}}$ and whose width tends to 0 in probability, and any fixed target c,*

$$\limsup_n \mathbb{P}(c \in I_n) \;\leq\; \mathbb{P}(L = c) \;\leq\; \max_x \mathbb{P}(L = x) < 1, \tag{11}$$

which is 0 whenever c is not an atom of L.

Corollary A.23 (The pooled number sits below the i.i.d. quantity at every finite design)**.** *Under Definition 3.1, $\mathbb{E}[\widehat{C}_{\text{pooled}}] \leq \text{pass@}k^{\text{iid}}$, with equality if and only if $\operatorname{Var}(\theta) = 0$, or $k = 1$, or $n = 1$. An allocation with every $c_r \leq 1$ contributes exactly $(1-\mu_1)^k$ to the product in* (10)*, that is, contributes $\text{pass@}k^{\text{iid}}$ to the expectation; when $\operatorname{Var}(\theta) > 0$, every allocation that puts two or more of the k chosen attempts inside one run contributes strictly more to the product and so strictly less to $\mathbb{E}[\widehat{C}_{\text{pooled}}]$. With Proposition 3.2 this places $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ between* succ@k *and* $\text{pass@}k^{\text{iid}}$.

Proposition A.24 (Cost-optimal number of continuations)**.** *Fix the total cost N, treat the number of runs $R = N/(c_0 + nc_1)$ as continuous (the large-budget relaxation) and write $\kappa := c_0/c_1$. Among designs of equal cost, minimising $\operatorname{Var}(\widehat{C}_{\text{run}})$ is equivalent to minimising*

$$\Phi(n) \;:=\; \big(V_0(\pi, k) + W_\pi(n, k)\big)\,(\kappa + n), \qquad n \in \{k, k+1, \dots\}, \tag{12}$$

and if $V_0(\pi, k) > 0$ the minimiser $n^\star$ is finite. Moreover: (i) as $\kappa \to 0$ the objective becomes $n\big(V_0 + W_\pi(n, k)\big)$, and $n^\star = k$ if and only if $V_0 \geq k\,W_\pi(k, k) - (k+1)\,W_\pi(k+1, k)$; at a general $\kappa \geq 0$ the criterion is (15)*. Both directions hold at every k: $n \mapsto n\,W_\pi(n, k)$ and Φ are convex by Lemma A.26; the closed form at $k = 2$ and the exact-arithmetic sweep for $k \in \{8, 16, 32\}$ on 505 mixing laws (Table S1) check the implementation rather than support the claim; (ii) if in addition $\pi((0,1)) > 0$ — equivalently $W_\pi(k, k) > 0$, which is what makes $n \mapsto W_\pi(n, k)$ strictly decreasing — then in this relaxation $n^\star \to \infty$ as $\kappa \to \infty$ (at a finite budget the feasible set $R \geq 1$ empties first,*

so the statement is about cost per unit of precision); without that condition the conclusion is false, a mixing law supported on $\{0,1\}$ *having* $W_\pi \equiv 0$, $\Phi(n) = V_0(\kappa + n)$ *and therefore* $n^\star = k$ *at every cost ratio. At a fixed number of runs* $\inf_n \mathrm{Var}(\widehat{C}_{\mathrm{run}}) = V_0/R$ *for every* π*, so no cost ratio buys away the floor; (iii)* $V_0 = 0$ *if and only if* π *is a point mass* δ_{θ_0}*, that is, only when there is no dependence to speak of; if in addition* $\theta_0 \in (0,1)$ *and either* $k \geq 2$ *or* $\kappa > 0$*, then* Φ *is strictly decreasing in* n *and* $n^\star \to \infty$ *— with no floor, the whole budget belongs inside a single run, which is the classical intuition and the exact condition under which it is right. The excluded case is a genuine exception: at* $k = 1$ *and* $\kappa = 0$*,* $W_\pi(n,1) = \theta_0(1-\theta_0)/n$ *exactly, so* $\Phi \equiv \theta_0(1-\theta_0)$ *and every* n *is optimal. At the two boundary masses* $\pi = \delta_0$ *and* $\pi = \delta_1$ *the claim degenerates rather than fails: every run is all failures or all successes, so* $W_\pi \equiv 0$*,* $\Phi \equiv 0$*, every design has variance zero and the allocation is indifferent.*

Proposition A.25 (Closed form of the within-run term)**.** $\widehat{C}_{n,k}$ *is the U-statistic with kernel* $h(x_1,\dots,x_k) = \max_i x_i$*, so for every* θ

$$\mathrm{Var}\big(\widehat{C}_{n,k}(X) \mid \theta\big) \;=\; \sum_{c=1}^{k} \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}}\, \zeta_c(\theta), \qquad \zeta_c(\theta) \;=\; (1-\theta)^{2k-c}\big(1-(1-\theta)^c\big), \tag{13}$$

and $n\,\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) \to k^2\zeta_1(\theta) = k^2\theta(1-\theta)^{2k-1}$*. The approach to that limit is at rate* $1/n$ *with a constant that depends on* k *and* θ*, so the asymptotic form should not be substituted for* (13) *at the values of* n *a budget actually permits.*

Lemma A.26 (Orthogonal form of the within-run term, and convexity at every k)**.** *Under Definition 3.1, for every mixing law* π*, every* $k \geq 1$ *and every* $n \geq k$*,*

$$W_\pi(n,k) \;=\; \sum_{j=1}^{k} \frac{a_j}{\binom{n}{j}}, \qquad a_j \;:=\; \binom{k}{j}^2 \mathbb{E}_\pi\big[\theta^j(1-\theta)^{2k-j}\big] \;\geq\; 0. \tag{14}$$

Consequently, on the real interval $n \geq k$*, both* $W_\pi(\cdot,k)$ *and* $n \mapsto n\,W_\pi(n,k)$ *are non-increasing and convex, and* $\Phi(n) = (\kappa+n)\big(V_0 + W_\pi(n,k)\big)$ *is* convex *for every* $\kappa \geq 0$ *and every constant* $V_0 \geq 0$*. Convexity is the only property of* Φ *used below, and it is the only one available:* Φ *need not be non-increasing, since at* $\pi = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ *and* $k = 2$ *one has* $W_\pi \equiv 0$ *and* $V_0 = \frac{1}{4}$*, so* $\Phi(n) = (\kappa+n)/4$ *is strictly increasing on* $n \geq 2$*. The decrease of* $n \mapsto n\,W_\pi(n,k)$ *is strict exactly when* $a_j > 0$ *for some* $j \geq 2$*, that is when* $k \geq 2$ *and* $\pi((0,1)) > 0$*. Hence, over the integers* $n \geq k$*,* $n^\star = k$ *minimises* Φ *if and only if*

$$V_0 + (\kappa + k + 1)\,W_\pi(k+1,k) \;-\; (\kappa+k)\,W_\pi(k,k) \;\geq\; 0. \tag{15}$$

Two limits are worth stating. It is a statement about the conditional-i.i.d. model of Definition 3.1, not about conditional exchangeability in general; and it concerns the relaxation, not the integer-feasible problem in which $R = \lfloor N/(\kappa+n) \rfloor$.

Table 4: At equal cost, crossing the threshold lowers the worst-case root mean squared error, except on a narrow interval around the headline design's mean, a restriction a measured dispersion does not supply. The last column, ratio, is the worst-case root mean squared error of the low-bias estimator on $R = k - 1$ runs divided by that of the unbiased $\widehat{U}_k$ on $R = k$ runs: above 1 the unbiased design has the smaller worst-case risk; at or below 1 (bold) the impossibility of unbiasedness is real but costs less than the variance the extra attempts buy back, which happens only on the headline family. The computation is analytical, with no experimental input beyond the two $(\hat{\rho}, \hat{\mu}_1)$ pairs and the cost ratios named below. The estimand is pass@k^{iid} on the two-point family π_α of (T), and the two designs have the same cost $R(c_0 + nc_1)$: $R = k$ runs of n_U attempts, where $\widehat{U}_k$ is unbiased, against $R = k - 1$ runs of n_L attempts, where Theorem 4.1 says no estimator is. On the second design the estimator is the one Theorem A.6 says attains the minimax bias: $L[p^\star] = \sum_{i \le R} a_i \, e_i(W) / \binom{R}{i}$ with $W_r = S_r/n$, e_i the ith elementary symmetric polynomial and a the coefficients of the best uniform degree-$(k-1)$ approximation $p^\star$ to $1 - (1 - q)^k$ on $[\theta_0, \theta_1]$. That family contains $\widehat{U}_k$: at $R \ge k$, L applied to the target itself is $\widehat{U}_k$ term for term (Remark A.7). Its worst-case bias reproduces the closed form $2(|\Delta|/4)^k$ of Theorem A.6(c) to nine significant figures and equioscillates. RMSE* is $\sup_\alpha \sqrt{\text{MSE}}$ over a 201-point grid of mixing weights; both moments are computed by exact enumeration of $\{0, \dots, n\}^R$, not by simulation. Families: *extreme* is Table 3's $(\theta_0, \theta_1) = (0, 1)$, on which μ_1 ranges over all of $[0, 1]$; *Arm A* and *headline* have spacing $\Delta = 2\sqrt{\rho \mu_1 (1 - \mu_1)}$ at the within-run correlations §6 measures, anchored inside $[0, 1]$. The bias of $L[p]$ depends on the mixing law only through μ_1 (Remark A.7), so the bias column of those two families is the worst case over every law whose mean is known to lie in $[\theta_0, \theta_1]$ — a restriction on μ_1 that a measured dispersion does not supply — and the extreme rows, $2 \cdot 4^{-k}$, are the minimax bias over classes containing the two-point family on $\{0, 1\}$, that is when nothing restricts the mixing law; a known bound on $\text{Var}(\theta)$ is a different restriction, under which estimators reading several attempts per run can do better. The run-start cost ratios $\kappa = c_0/c_1$ are Table 37's amortised and fitted estimates, and zero. n_L is rounded down, so the low-bias design never costs more than the unbiased one. On the two restricted families the worst-case bias of the low-bias estimator is 4.7 to 277 times smaller than its own worst-case root mean squared error (4.7 for the headline design at $k = 2$, 277 for the headline design at $k = 4$), and only a few times smaller at $k = 2$, where it is 1.03×10^{-2} against a worst-case root mean squared error of 0.0479.

Family	k	κ	unbiased, $R = k$		low-bias, $R = k - 1$			ratio
			(R, n)	RMSE*	(R, n)	$\sup_\alpha \lvert\text{bias}\rvert$	RMSE*	
extreme: $(\theta_0, \theta_1) = (0.0000, 1.0000)$								
extreme	2	0.00	$(2, 4)$	0.5000	$(1, 8)$	1.25×10^{-1}	0.5154	1.031
extreme	2	1.15	$(2, 4)$	0.5000	$(1, 9)$	1.25×10^{-1}	0.5154	1.031
extreme	2	2.96	$(2, 4)$	0.5000	$(1, 10)$	1.25×10^{-1}	0.5154	1.031
extreme	2	0.00	$(2, 8)$	0.5000	$(1, 16)$	1.25×10^{-1}	0.5154	1.031
extreme	2	1.15	$(2, 8)$	0.5000	$(1, 17)$	1.25×10^{-1}	0.5154	1.031
extreme	2	2.96	$(2, 8)$	0.5000	$(1, 18)$	1.25×10^{-1}	0.5154	1.031
extreme	3	0.00	$(3, 4)$	0.5000	$(2, 6)$	3.13×10^{-2}	0.5913	1.183
extreme	3	1.15	$(3, 4)$	0.5000	$(2, 6)$	3.13×10^{-2}	0.5913	1.183
extreme	3	2.96	$(3, 4)$	0.5000	$(2, 7)$	3.13×10^{-2}	0.5913	1.183
extreme	3	0.00	$(3, 8)$	0.5000	$(2, 12)$	3.13×10^{-2}	0.5913	1.183
extreme	3	1.15	$(3, 8)$	0.5000	$(2, 12)$	3.13×10^{-2}	0.5913	1.183
extreme	3	2.96	$(3, 8)$	0.5000	$(2, 13)$	3.13×10^{-2}	0.5913	1.183
extreme	4	0.00	$(4, 4)$	0.5000	$(3, 5)$	7.81×10^{-3}	0.5977	1.195
extreme	4	1.15	$(4, 4)$	0.5000	$(3, 5)$	7.81×10^{-3}	0.5977	1.195
extreme	4	2.96	$(4, 4)$	0.5000	$(3, 6)$	7.81×10^{-3}	0.5977	1.195
extreme	4	0.00	$(4, 8)$	0.5000	$(3, 10)$	7.81×10^{-3}	0.5977	1.195
extreme	4	1.15	$(4, 8)$	0.5000	$(3, 11)$	7.81×10^{-3}	0.5977	1.195
extreme	4	2.96	$(4, 8)$	0.5000	$(3, 11)$	7.81×10^{-3}	0.5977	1.195
Arm A: $(\theta_0, \theta_1) = (0.1310, 0.9430)$								
Arm A	2	0.00	$(2, 4)$	0.3927	$(1, 8)$	8.24×10^{-2}	0.3964	1.010
Arm A	2	1.15	$(2, 4)$	0.3927	$(1, 9)$	8.24×10^{-2}	0.3951	1.006
Arm A	2	2.96	$(2, 4)$	0.3927	$(1, 10)$	8.24×10^{-2}	0.3941	1.004
Arm A	2	0.00	$(2, 8)$	0.3749	$(1, 16)$	8.24×10^{-2}	0.3907	1.042
Arm A	2	1.15	$(2, 8)$	0.3749	$(1, 17)$	8.24×10^{-2}	0.3903	1.041

continued on the next page

Table 4, continued.

Family	k	κ	unbiased, $R = k$		low-bias, $R = k - 1$			ratio
			(R, n)	RMSE*	(R, n)	$\sup_\alpha \lvert\text{bias}\rvert$	RMSE*	
Arm A	2	2.96	$(2, 8)$	0.3749	$(1, 18)$	8.24×10^{-2}	0.3900	1.041
Arm A	3	0.00	$(3, 4)$	0.3555	$(2, 6)$	1.67×10^{-2}	0.4030	1.133
Arm A	3	1.15	$(3, 4)$	0.3555	$(2, 6)$	1.67×10^{-2}	0.4030	1.133
Arm A	3	2.96	$(3, 4)$	0.3555	$(2, 7)$	1.67×10^{-2}	0.3990	1.122
Arm A	3	0.00	$(3, 8)$	0.3332	$(2, 12)$	1.67×10^{-2}	0.3888	1.167
Arm A	3	1.15	$(3, 8)$	0.3332	$(2, 12)$	1.67×10^{-2}	0.3888	1.167
Arm A	3	2.96	$(3, 8)$	0.3332	$(2, 13)$	1.67×10^{-2}	0.3877	1.164
Arm A	4	0.00	$(4, 4)$	0.3215	$(3, 5)$	3.40×10^{-3}	0.3691	1.148
Arm A	4	1.15	$(4, 4)$	0.3215	$(3, 5)$	3.40×10^{-3}	0.3691	1.148
Arm A	4	2.96	$(4, 4)$	0.3215	$(3, 6)$	3.40×10^{-3}	0.3616	1.125
Arm A	4	0.00	$(4, 8)$	0.2957	$(3, 10)$	3.40×10^{-3}	0.3467	1.172
Arm A	4	1.15	$(4, 8)$	0.2957	$(3, 11)$	3.40×10^{-3}	0.3447	1.166
Arm A	4	2.96	$(4, 8)$	0.2957	$(3, 11)$	3.40×10^{-3}	0.3447	1.166
headline: $(\theta_0, \theta_1) = (0.7131, 1.0000)$								
headline	2	0.00	$(2, 4)$	0.1050	$(1, 8)$	1.03×10^{-2}	0.0546	**0.520**
headline	2	1.15	$(2, 4)$	0.1050	$(1, 9)$	1.03×10^{-2}	0.0533	**0.507**
headline	2	2.96	$(2, 4)$	0.1050	$(1, 10)$	1.03×10^{-2}	0.0522	**0.497**
headline	2	0.00	$(2, 8)$	0.0707	$(1, 16)$	1.03×10^{-2}	0.0486	**0.687**
headline	2	1.15	$(2, 8)$	0.0707	$(1, 17)$	1.03×10^{-2}	0.0482	**0.682**
headline	2	2.96	$(2, 8)$	0.0707	$(1, 18)$	1.03×10^{-2}	0.0479	**0.677**
headline	3	0.00	$(3, 4)$	0.0426	$(2, 6)$	7.38×10^{-4}	0.0300	**0.704**
headline	3	1.15	$(3, 4)$	0.0426	$(2, 6)$	7.38×10^{-4}	0.0300	**0.704**
headline	3	2.96	$(3, 4)$	0.0426	$(2, 7)$	7.38×10^{-4}	0.0274	**0.642**
headline	3	0.00	$(3, 8)$	0.0264	$(2, 12)$	7.38×10^{-4}	0.0212	**0.804**
headline	3	1.15	$(3, 8)$	0.0264	$(2, 12)$	7.38×10^{-4}	0.0212	**0.804**
headline	3	2.96	$(3, 8)$	0.0264	$(2, 13)$	7.38×10^{-4}	0.0206	**0.781**
headline	4	0.00	$(4, 4)$	0.0165	$(3, 5)$	5.29×10^{-5}	0.0147	**0.890**
headline	4	1.15	$(4, 4)$	0.0165	$(3, 5)$	5.29×10^{-5}	0.0147	**0.890**
headline	4	2.96	$(4, 4)$	0.0165	$(3, 6)$	5.29×10^{-5}	0.0127	**0.771**
headline	4	0.00	$(4, 8)$	0.0095	$(3, 10)$	5.29×10^{-5}	0.0088	**0.931**
headline	4	1.15	$(4, 8)$	0.0095	$(3, 11)$	5.29×10^{-5}	0.0083	**0.874**
headline	4	2.96	$(4, 8)$	0.0095	$(3, 11)$	5.29×10^{-5}	0.0083	**0.874**

A.2 Proofs for Section 3

This subsection proves the results of Section 3 on the model, the three quantities and uniqueness, as stated in §A.1.

Proof of Lemma A.3. Let $\bar{T}(x) := c_{|x|}$ be the average of T over the orbit of x under coordinate permutations. Under $\pi = \delta_\theta$ the law $\text{Bern}(\theta)^{\otimes n}$ is permutation invariant, so

$$\mathbb{E}_{\delta_\theta}[T] \;=\; \mathbb{E}_{\delta_\theta}[\bar{T}] \;=\; \sum_{s=0}^{n} c_s \binom{n}{s} \theta^s (1-\theta)^{n-s} \;=\; B_c(\theta),$$

and conditioning on θ and integrating against π gives $\mathbb{E}_\pi[T] = \mathbb{E}_\pi[B_c(\theta)]$ for every π.

The $n + 1$ Bernstein polynomials $\binom{n}{s}\theta^s(1-\theta)^{n-s}$ are linearly independent: the substitution $\theta = u/(1+u)$ followed by multiplication by $(1+u)^n$ turns $B_c \equiv 0$ into $\sum_s c_s \binom{n}{s} u^s \equiv 0$ for $u > 0$, which forces $c = 0$. They are therefore $n + 1$ independent elements of the $(n + 1)$-dimensional space of polynomials of degree at most n, hence a basis, and $c \mapsto B_c$ is a linear bijection. Expanding B_c in monomials and integrating term by term identifies $\{\pi \mapsto \mathbb{E}_\pi[T] : T\}$ with $\{\pi \mapsto \sum_{j=0}^{n} a_j \mu_j\}$, the coefficient vectors a and c being in linear bijection.

No hypothesis on Π has been used. The independent-attempt target is outside this set for $k \geq 2$ *when* Π *satisfies the richness condition* (T), and that hypothesis is needed rather than decorative: on $\Pi = \{\delta_p : p \in [0,1]\}$ with $n \geq k$ the independent-attempt target is $1-(1-p)^k$, a polynomial of degree $k \leq n$ in p and therefore a linear functional of the first n moments — one that $\widehat{C}_{n,k}$ estimates without bias. So assume (T), which supplies a segment $\pi_\alpha = (1-\alpha)\delta_{\theta_0} + \alpha\delta_{\theta_1}$, α ranging over an infinite $A \subseteq [0,1]$, inside Π. Along it every moment $\mu_j(\pi_\alpha)$ is affine in α, so every functional in the image is affine in α, whereas $1-(1-\mu_1(\pi_\alpha))^k$ is a polynomial in α of degree exactly k; two polynomials that agree on an infinite set agree identically, so no member of the image can equal it. Hence pass@k^{iid} is not an affine functional of the moment sequence, while succ@k is, by (2). □

Proof of Proposition 3.3. Let S be a uniformly random k-subset of $[n]$, drawn independently of X. Given X with c successes, the k-subsets that avoid every success are exactly the k-subsets of the $n-c$ failed positions, so $\mathbb{P}(\forall i \in S : X_i = 0 \mid X) = \binom{n-c}{k}/\binom{n}{k}$ and therefore $\widehat{C}_{n,k}(X) = \mathbb{P}(\exists i \in S : X_i = 1 \mid X)$; the convention $\binom{m}{k} = 0$ for $m < k$ is what makes this hold when the failures are too few to fill a subset. Taking expectations and using $S \perp X$,

$$\mathbb{E}_G[\widehat{C}_{n,k}] \;=\; \binom{n}{k}^{-1} \sum_{|S|=k} \mathbb{P}_G(\exists i \in S : X_i = 1),$$

which is the subset-averaged success probability. If G is exchangeable then for each $S = \{i_1 < \cdots < i_k\}$ the vector $(X_{i_1}, \ldots, X_{i_k})$ has the law of $(X_1, \ldots, X_k)$, so every summand equals $\mathbb{P}_G(X_1 \vee \cdots \vee X_k = 1)$, and so does their average. Equivalently: $\sum_{|S|=k} \mathbf{1}\{\exists i \in S : x_i = 1\} = \binom{n}{k} - \binom{n-c}{k}$ exhibits $\widehat{C}_{n,k}$ as the U-statistic with kernel $h(x_1, \ldots, x_k) = \max_i x_i$, and the unbiasedness of a U-statistic for the expectation of its kernel under an exchangeable law is classical. □

Proof of Corollary A.4. Sufficiency: $P_\pi(X = x)$ depends on x only through $|x|$, so the conditional law of X given $S = s$ is uniform on the orbit $\{|x| = s\}$ and does not involve π. Completeness: suppose $\mathbb{E}_\pi[g(S)] = 0$ for every $\pi \in \Pi$. Specialising to $\pi = \delta_\theta$ gives $B_g(\theta) = 0$ for every $\theta \in \Theta$; a polynomial vanishing on an infinite set is identically zero, so $B_g \equiv 0$ and $g \equiv 0$ by the bijection of Lemma A.3. Since $\widehat{C}_{n,k}$ is a function of S and is unbiased for succ@k by Proposition 3.3, Lehmann–Scheffé gives that it is the UMVUE. □

Proof of Proposition 3.4. $T = \mathbf{1}\{X_1 \vee \cdots \vee X_k = 1\}$ is the indicator of the event whose probability is the estimand, so $\mathbb{E}_G[T] = \mathrm{succ@}k(G)$ by definition, and T is $\{0,1\}$-valued, so $\mathrm{Var}_G(T) = \mathrm{succ@}k(1 - \mathrm{succ@}k)$. The argument uses neither exchangeability nor a latent-variable representation, and restricts neither the dependence between attempts nor the adaptivity of the procedure; it uses only that the first k attempts are observed. □

Proof of Proposition 3.2. Write $Z := 1-\theta \in [0,1]$ and $m_c := \mathbb{E}_\pi[Z^c]$, so that succ@$k = 1 - m_k$ and pass@$k^{\mathrm{iid}} = 1 - m_1^k$, and by Proposition A.21 $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] = 1 - \mathbb{E}_c\left[\prod_r m_{c_r}\right]$ over the hypergeometric allocation c with $\sum_r c_r = k$. *Upper bound.* By Jensen, $m_c \geq m_1^c$ for every c, so $\prod_r m_{c_r} \geq m_1^k$ and $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] \leq \mathrm{pass@}k^{\mathrm{iid}}$; the equality cases are those of Corollary A.23. *Lower bound.* Lyapunov's inequality, $m_c^{1/c} \leq m_k^{1/k}$ for $1 \leq c \leq k$, gives $m_c \leq m_k^{c/k}$, and $m_0 = 1 = m_k^0$; hence $\prod_r m_{c_r} \leq m_k^{\sum_r c_r/k} = m_k$ for every allocation, and $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] \geq 1 - m_k = \mathrm{succ@}k$. *Equality.* If $\mathrm{Var}(\theta) = 0$ all three quantities coincide; if $k = 1$ every allocation is a single attempt; if $R = 1$ every allocation is $c = (k)$, whose product is m_k. Conversely let $\mathrm{Var}(\theta) > 0$, $k \geq 2$ and $R \geq 2$. Lyapunov's inequality is strict for $0 < c < k$ unless Z is almost surely constant, so $m_c < m_k^{c/k}$ there, and $m_c > 0$ for every c because Z is not almost surely 0. Since $nR \geq k$, the allocation that places $c_1 = \min(n, k-1) \in [1, k-1]$ attempts in the first run and the remaining $k - c_1 \leq n(R-1)$ in the others has positive probability, and its product is strictly below m_k; so $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] > \mathrm{succ@}k$. The same argument with Jensen's inequality, strict for $c \geq 2$, gives the strictness condition of the upper bound: an allocation with some $c_r \geq 2$ has positive probability exactly when $n \geq 2$ and $k \geq 2$. Finally, a given pair of the k selected attempts lies in one run with probability $(n-1)/(nR-1)$, so some run receives two of them with probability at most $\binom{k}{2}(n-1)/(nR-1) \to 0$ as $R \to \infty$ at fixed

n and k; every other allocation contributes $(1-\mu_1)^k$, so $\mathbb{E}[\widehat{C}_{\text{pooled}}] \to \text{pass@}k^{\text{iid}}$, while at $R=1$ it equals $\text{succ@}k$ (Proposition 3.3). □

Remark A.27 (Scope of Proposition 3.2). The ordering uses the conditionally i.i.d. structure of Definition 3.1; exchangeability within a run is not enough. With $n=k=3$ and $R=2$, let each run independently consist, with probability $1/2$ each, of three successes or of exactly one success at a uniformly random position. The within-run law is exchangeable, $\text{succ@}3 = 1$ and $\mu_1 = 2/3$, so $\text{pass@}3^{\text{iid}} = 26/27 \approx 0.963$; the pooled estimator falls below one only when both runs have a single success, where it equals $1 - \binom{4}{3}/\binom{6}{3} = 0.8$, so $\mathbb{E}[\widehat{C}_{\text{pooled}}] = 0.95$, below both estimands. On the ordered tree-search design the pooled report also sits above the first-attempt $\widehat{U}_4$ (Table 50).

A.3 Proofs for Section 4

This subsection proves the results of Section 4 on the run-count threshold and what it costs below it, as stated in §A.1.

Proof of Theorem 4.1(a). ($\Leftarrow$) Let $R \geq k$. By θ-calibration, the first attempt $X_1^{(r)}$ of run r is a measurable function of O_r with $\mathbb{P}(X_1^{(r)} = 1 \mid \theta_r) = \theta_r$, so $\mathbb{P}(X_1^{(r)} = 0) = \mathbb{E}_\pi[1-\theta] = 1-\mu_1$. Runs being independent, $T := \mathbf{1}\{\exists r \leq k : X_1^{(r)} = 1\}$ has $\mathbb{E}[T] = 1-(1-\mu_1)^k = \text{pass@}k^{\text{iid}}$ for every $\pi \in \Pi$. Proposition A.8 is the symmetrised version of the same estimator.

($\Rightarrow$) Let $R < k$ and suppose $T : O^R \to \mathbb{R}$ is unbiased for $\text{pass@}k^{\text{iid}}$ on all of Π. Restrict π to the two-point family π_α, $\alpha \in A$, supplied by richness (T). The law of one run's observation is $\int K(\theta,\cdot)\,\pi_\alpha(d\theta) = (1-\alpha)K(\theta_0,\cdot) + \alpha K(\theta_1,\cdot)$, affine in α, and the joint law of the R independent runs is its R-fold product,

$$P_\alpha^{(R)} \;=\; \sum_{c\in\{0,1\}^R} (1-\alpha)^{R-|c|}\alpha^{|c|} \bigotimes_{r=1}^{R} K(\theta_{c_r},\cdot), \tag{16}$$

a mixture of 2^R fixed measures whose weights are monomials in α of degree at most R. Consequently

$$\varphi_T(\alpha) \;:=\; \mathbb{E}_{\pi_\alpha}[T] \;=\; \sum_{c\in\{0,1\}^R} (1-\alpha)^{R-|c|}\alpha^{|c|} A_c, \qquad A_c := \int T\, d\Big(\bigotimes_r K(\theta_{c_r},\cdot)\Big),$$

so φ_T is a polynomial in α of degree at most R — provided every A_c is finite and the interchange is legitimate. It is: A is infinite, so it meets $(0,1)$; fix $\alpha^* \in A \cap (0,1)$. Unbiasedness at α^* presupposes $\mathbb{E}_{\pi_{\alpha^*}}|T| < \infty$, and applying (16) to $|T| \geq 0$ writes that finite number as a sum of non-negative terms whose weights $(1-\alpha^*)^{R-|c|}(\alpha^*)^{|c|}$ are all strictly positive; hence $\int |T|\, d(\bigotimes_r K(\theta_{c_r},\cdot)) < \infty$ for every c, each A_c converges absolutely, and Fubini applies at every α.

Along the same family, $\mu_1(\pi_\alpha) = \theta_0 + \alpha\Delta$ with $\Delta = \theta_1 - \theta_0 \neq 0$, so $\psi(\alpha) := \text{pass@}k^{\text{iid}}(\pi_\alpha) = 1-(1-\theta_0-\alpha\Delta)^k$ has coefficient $(-1)^{k+1}\Delta^k \neq 0$ on α^k and therefore degree exactly k. Unbiasedness gives $\varphi_T = \psi$ on the infinite set A; two polynomials agreeing on an infinite set are equal, so $k = \deg\psi = \deg\varphi_T \leq R < k$, a contradiction.

External randomisation changes nothing. If $T = T(O,U)$ with $U \sim \nu$ independent of O, then for each u the coefficients $A_c(u) = \int T(\cdot,u)\, d(\bigotimes_r K(\theta_{c_r},\cdot))$ are defined by the same argument applied to the joint integrability of $|T|$ at α^*, and

$$\mathbb{E}_{\pi_\alpha}[T] \;=\; \int \varphi_{T(\cdot,u)}(\alpha)\,\nu(du) \;=\; \sum_c (1-\alpha)^{R-|c|}\alpha^{|c|} \int A_c(u)\,\nu(du)$$

is again a polynomial of degree at most R; the degree count is unchanged.

Nothing in the argument refers to n, to the structure of O, or to any property of K beyond its not depending on π, which is why enriching the observation cannot help. □

Proof of Corollaries A.11, A.12, A.5 and A.13. Corollary A.11. $R = 1 < k$ for every $k \geq 2$; apply Theorem 4.1.

Corollary A.12. Unbiasedness on the i.i.d. submodel reads $B_c(\theta) = 1 - (1-\theta)^k$ for every $\theta \in \Theta$, hence as an identity of polynomials. With the convention $\binom{n-k}{s} = 0$ for $s > n-k$,

$$\sum_{s=0}^{n} \binom{n-k}{s} \theta^s (1-\theta)^{n-s} = (1-\theta)^k \sum_{s} \binom{n-k}{s} \theta^s (1-\theta)^{(n-k)-s} = (1-\theta)^k,$$

so the Bernstein coefficients of $(1-\theta)^k$ are $\binom{n-k}{s}/\binom{n}{s}$, and uniqueness of Bernstein coefficients gives $c_s = 1 - \binom{n-k}{s}/\binom{n}{s}$. Both sides of $\binom{n-k}{s}\binom{n}{k} = \binom{n-s}{k}\binom{n}{s}$ equal $n!/\big(s!\,k!\,(n-s-k)!\big)$ when $s+k \le n$, and both vanish otherwise, so $c_s = 1 - \binom{n-s}{k}/\binom{n}{k}$, which is (3) evaluated at $c = s$. For part (b): an exchangeable G satisfies $\mathbb{E}_G[T] = \mathbb{E}_G[\bar{T}]$ and $\bar{T}(x) = c_{|x|} = \widehat{C}_{n,k}(x)$, so Proposition 3.3 gives $\mathbb{E}_G[T] = \mathrm{succ}@k(G)$.

Corollary A.5. Write $\bar{u} = 1 - \mu_1$ and expand $(1-\theta)^k = \big(\bar{u} - (\theta - \mu_1)\big)^k$ by the binomial theorem; taking expectations and using $\gamma_0 = 1$, $\gamma_1 = 0$,

$$\mathrm{pass}@k^{\mathrm{iid}} - \mathrm{succ}@k \;=\; \mathbb{E}\big[(1-\theta)^k\big] - \bar{u}^{\,k} \;=\; \sum_{j=2}^{k} \binom{k}{j} \bar{u}^{\,k-j} (-1)^j \gamma_j,$$

an exact finite identity rather than an approximation. Since $x \mapsto (1-x)^k$ is strictly convex on $[0,1]$ for $k \ge 2$, Jensen gives $\mathbb{E}[(1-\theta)^k] \ge (1-\mu_1)^k$ with equality exactly when θ is almost surely constant, i.e. $\mathrm{Var}(\theta) = 0$; this is the stated sign. The $j = 2$ term is $\binom{k}{2}(1-\mu_1)^{k-2}\,\mathrm{Var}(\theta)$, and $\mathrm{Cov}(X_1, X_2) = \mathbb{E}[\theta^2] - \mu_1^2 = \mathrm{Var}(\theta)$ with $\mathrm{Var}(X_1) = \mu_1(1-\mu_1)$ gives $\mathrm{Var}(\theta) = \rho\,\mu_1(1-\mu_1)$. For the remainder, $|\theta - \mu_1| \le 1$ implies $|\gamma_j| \le \mathrm{Var}(\theta)$ for $j \ge 2$, so $\big|\sum_{j\ge 3}\big| \le \mathrm{Var}(\theta) \sum_{j=3}^{k} \binom{k}{j} \bar{u}^{\,k-j} = \mathrm{Var}(\theta)\big[(1+\bar{u})^k - \bar{u}^k - k\bar{u}^{k-1} - \binom{k}{2}\bar{u}^{k-2}\big]$.

Corollary A.13. The product structure is what lets question q^* move while every other π_q stays fixed. Perturb only question q^* along its two-point family. Since $R_{q^*} = 1$ and the questions are independent, the joint law of the whole observation is affine in α, so $\alpha \mapsto \mathbb{E}[T]$ has degree at most 1, while the target $Q^{-1}\big[1-(1-\theta_0-\alpha\Delta)^k\big]+\mathrm{const}$ has degree exactly $k \ge 2$. The same argument applies whenever some question has $R_q < k$. □

Proof of Theorem A.6. (a) Put $\theta_0 = a$, $\theta_1 = b$ and $\Delta = b - a$ in the proof of Theorem 4.1: for every estimator T, randomised or not, $\alpha \mapsto \mathbb{E}_{\pi_\alpha}[T]$ is the restriction to the weight set of some polynomial of degree at most R in α, hence of some $p \in \mathcal{P}_R$ in $q = a + \alpha\Delta$, while $\mathrm{pass}@k^{\mathrm{iid}}(\pi_\alpha) = \psi(q)$. For the continuous function $p - \psi$ the supremum over a dense set of weights equals the supremum over $q \in I$, so the worst-case bias is at least $\max_I |p - \psi| \ge d_R(I)$. Density is needed: with $k = 2$, $R = 1$, $I = [0,1]$ and weights $\{\varepsilon/j : j \ge 1\}$, an infinite set accumulating only at 0, the estimator $T = 2X_1^{(1)}$ has $\mathbb{E}_{\pi_\alpha}[T] = 2\alpha$ against $\psi(\alpha) = 2\alpha - \alpha^2$, so its worst-case bias over the family is ε^2, while $d_1([0,1]) = 1/8$.

(b) By θ-calibration each O_r carries the first-attempt outcome $B_r := X_1^{(r)}$ with $\mathbb{P}(B_r = 1 \mid \theta_r) = \theta_r$; runs being independent, the B_r are i.i.d. $\mathrm{Bern}(\mu_1)$ under *every* law of the class. For $T = f(B_1, \dots, B_R)$,

$$\mathbb{E}_\pi[T] \;=\; \sum_{j=0}^{R} \bar{f}_j \binom{R}{j} \mu_1^j (1-\mu_1)^{R-j}, \qquad \bar{f}_j := \binom{R}{j}^{-1} \sum_{b \in \{0,1\}^R,\ |b| = j} f(b),$$

and the $\bar{f}_j$ can be given any prescribed values, so as f varies $\mathbb{E}_\pi[T]$ ranges over all of $\mathcal{P}_R$ evaluated at μ_1. Choosing f with $\mathbb{E}_\pi[T] = p^\star(\mu_1)$ gives a bias $p^\star(\mu_1) - \psi(\mu_1)$ that depends on π only through μ_1 and is at most $d_R(I)$ in absolute value whenever $\mu_1 \in I$. With (a) this is the stated equality.

(c) ψ has degree k with leading coefficient $(-1)^{k+1}$, and the least sup-norm on an interval of length $b - a$ of a monic polynomial of degree k is $2\big((b-a)/4\big)^k$, attained by the rescaled Chebyshev polynomial, so $d_{k-1}(I) = 2\big((b-a)/4\big)^k$, with the error curve equioscillating at $k+1$ points. For $R = 1$ on $[0,1]$, ψ is increasing and strictly concave; the best uniform affine approximation of a concave function is its chord displaced by half the maximal gap. The chord is $q \mapsto q$, and the gap $\psi(q) - q$ is maximised where $k(1-q)^{k-1} = 1$, i.e. $1 - q = k^{-1/(k-1)}$, giving the gap $k^{-1/(k-1)} - k^{-k/(k-1)}$; the error of the optimal line equioscillates at three points, which certifies optimality. □

Proof of Theorem 4.1(b). Let $\pi, \pi' \in \Pi$ be mutually absolutely continuous with $\mu_1(\pi) \neq \mu_1(\pi')$ — under (T), π_α and $\pi_{\alpha'}$ for two distinct $\alpha, \alpha' \in A \cap (0,1)$, which share the support $\{\theta_0, \theta_1\}$; since $t \mapsto 1-(1-t)^k$ is injective on $[0,1]$, the target values $\psi(\pi)$ and $\psi(\pi')$ differ. Write P_n^π for the law of the R-run observation when n attempts per run are recorded. Both P_n^π and $P_n^{\pi'}$ are images of the latent laws $\pi^{\otimes R}$ and $\pi'^{\otimes R}$ under one and the same Markov kernel — the kernel generating attempts from the latent vector, which does not depend on the mixing law — so the data-processing inequality for total variation gives

$$\mathrm{TV}\big(P_n^\pi, P_n^{\pi'}\big) \;\le\; \mathrm{TV}\big(\pi^{\otimes R}, \pi'^{\otimes R}\big) \;=:\; 1-2\beta \;<\; 1$$

for *every* n, the strict inequality because mutual absolute continuity passes to the product measures. Fix $\varepsilon < |\psi(\pi)-\psi(\pi')|/2$ and let $A_n := \{|T_n - \psi(\pi)| < \varepsilon\}$, which is disjoint from $\{|T_n - \psi(\pi')| < \varepsilon\}$. Le Cam's two-point inequality gives

$$P_n^\pi(A_n^c) + P_n^{\pi'}(A_n) \;\ge\; 1 - \mathrm{TV}\big(P_n^\pi, P_n^{\pi'}\big) \;\ge\; 2\beta \;>\; 0$$

uniformly in n, whereas a sequence consistent under both laws would send both terms to zero. The bound holds for every fixed R — including $R \ge k$, where unbiased estimation is available but consistency in n is not — and for $k = 1$ as well. The same bound holds for $\mathrm{succ}@k$, since $\alpha \mapsto \mathrm{succ}@k(\pi_\alpha)$ is affine with non-zero slope, so distinct α give distinct targets.

The mechanism is visible at $R = 1$: the empirical attempt rate converges to θ, not to μ_1, so any plug-in converges to the random variable $1-(1-\theta)^k$, whose mean is $\mathrm{succ}@k$. Further attempts measure the same draw of θ more precisely; consistency requires $R \to \infty$. □

Proofs of Propositions A.8, A.14 and A.9. Proposition A.8. For distinct indices $r_1, \dots, r_k$ the runs are independent and, every attempt having success probability p, $\mathbb{E}[1 - S_r/n] = 1-p$ whatever the dependence inside a run, so $\mathbb{E}\big[\prod_{i\le k}(1 - S_{r_i}/n)\big] = (1-p)^k$; averaging over the ordered k-tuples of distinct indices leaves the value unchanged, whence $\mathbb{E}[\widehat{U}_k] = 1-(1-p)^k$.

Proposition A.14. Write $b_s(\pi) := P_\pi(S_1 = s) = \mathbb{E}_\pi\big[\binom{n}{s}\theta^s(1-\theta)^{n-s}\big]$. Given $S_r = s_r$ the vector $X^{(r)}$ is uniform on its orbit, so

$$P_\pi\big(x^{(1)}, \dots, x^{(R)}\big) = \prod_r \frac{b_{s_r}(\pi)}{\binom{n}{s_r}} = \Big[\prod_s b_s(\pi)^{\mathbf{N}_s}\Big] \cdot \Big[\prod_r \binom{n}{s_r}\Big]^{-1},$$

whose second factor is free of π; by the factorisation criterion $\mathbf{N}$ is sufficient. For completeness, suppose $\mathbb{E}_\pi[g(\mathbf{N})] = 0$ for all $\pi \in \Pi$. Then $\mathbb{E}_\pi[g(\mathbf{N})] = \sum_{\mathbf{N}} g(\mathbf{N}) \frac{R!}{\prod_s \mathbf{N}_s!} \prod_s b_s(\pi)^{\mathbf{N}_s} =: P_g\big(b(\pi)\big)$ with P_g a homogeneous polynomial of degree R in $(b_0, \dots, b_n)$. The map $(1, \mu_1, \dots, \mu_n) \mapsto (b_0, \dots, b_n)$ is an invertible linear map — the change of basis between monomials and the Bernstein basis of Lemma A.3 — so a moment image with non-empty interior in $\mathbb{R}^n$ produces a set $U \subseteq \{b(\pi) : \pi \in \Pi\}$ that is relatively open in the hyperplane $H = \{\sum_s b_s = 1\}$. Now P_g vanishes on U and is homogeneous, so it vanishes on the open cone $\{t\,u : t > 0,\ u \in U\}$, a non-empty open subset of $\mathbb{R}^{n+1}$; a polynomial vanishing on a non-empty open set is identically zero. Distinct multisets $\mathbf{N}$ give distinct monomials $\prod_s b_s^{\mathbf{N}_s}$, so every coefficient $g(\mathbf{N})\,R!/\prod_s \mathbf{N}_s!$ vanishes and $g \equiv 0$. Both $\widehat{U}_k$ and $R^{-1}\sum_r \widehat{C}_{n,k}(X^{(r)})$ are symmetric functions of $(S_1, \dots, S_R)$, hence $\mathbf{N}$-measurable, and they are unbiased for $\mathrm{pass}@k^{\mathrm{iid}}$ and for $\mathrm{succ}@k$ respectively; Lehmann–Scheffé gives (i), (ii) and (iii).

Proposition A.9. The first inequality is Proposition A.14(ii). For the second, condition on $\theta = (\theta_1, \dots, \theta_R)$: the runs are conditionally independent with $\mathbb{E}[1 - S_r/n \mid \theta] = 1-\theta_r$, so $\mathbb{E}[\widehat{U}_k \mid \theta] = U_k^\theta$ and the law of total variance gives $\mathrm{Var}(\widehat{U}_k) = \mathrm{Var}(U_k^\theta) + \mathbb{E}[\mathrm{Var}(\widehat{U}_k \mid \theta)] \ge \mathrm{Var}(U_k^\theta)$. For the third, U_k^θ is a U-statistic of order k in the i.i.d. sample $\theta_1, \dots, \theta_R$ with symmetric kernel $h(t_1, \dots, t_k) = 1 - \prod_i (1-t_i)$. Its Hájek projection onto sums of functions of a single θ_r satisfies $\mathbb{E}[U_k^\theta \mid \theta_1] - \mathbb{E}[U_k^\theta] = \frac{\binom{R-1}{k-1}}{\binom{R}{k}}\, g_1(\theta_1) = \frac{k}{R} g_1(\theta_1)$ with $g_1(t) := \mathbb{E}[h(t, \theta_2, \dots, \theta_k)] - \mathbb{E}[h]$, so the projection has variance $R\,(k/R)^2\zeta_1 = (k^2/R)\zeta_1$, and a projection in L^2 never has larger variance than the projected variable (Hoeffding, 1948). Finally $\mathbb{E}[h(\theta_1, \dots, \theta_k) \mid \theta_1] = 1-(1-\theta_1)(1-\mu_1)^{k-1}$ gives $\zeta_1 = (1-\mu_1)^{2(k-1)}\,\mathrm{Var}(\theta)$. The resulting bound involves neither n nor the within-run structure. □

Proofs of Propositions A.15, A.17 and Theorem 3.5. Proposition A.15. Write $u := 1-\theta$. Under δ_θ the stopped observation $W = (\tau, X_\tau)$ takes the $N_{\max}+1$ values $(j, 1)$ for $1 \le j \le N_{\max}$ and $(N_{\max}, 0)$,

with probabilities $u^{j-1}(1-u)$ and $u^{N_{\max}}$. Hence for any $t : W \mapsto \mathbb{R}$,

$$\mathbb{E}_{\delta_\theta}[t(W)] \;=\; \sum_{j=1}^{N_{\max}} t(j,1)\, u^{j-1}(1-u) + t(N_{\max},0)\, u^{N_{\max}} \;=:\; h(u).$$

The $N_{\max}+1$ functions $\{u^{j-1}(1-u)\}_{j\le N_{\max}} \cup \{u^{N_{\max}}\}$ span polynomials of degree at most $N_{\max}$: telescoping gives $\sum_{j\le m} u^{j-1}(1-u) = 1-u^m$ for $m = 1,\dots,N_{\max}$, adding $u^{N_{\max}}$ to $1-u^{N_{\max}}$ gives the constant 1, hence every u^m with $m \le N_{\max}$ is in the span. Since both spaces have dimension $N_{\max}+1$, the coefficient map $t \mapsto h$ is a linear bijection. Composing with Lemma A.3 shows the estimable functionals from W are exactly those estimable from the full attempt vector. Since Π contains δ_θ for infinitely many θ, a polynomial of degree at most $N_{\max}$ vanishing at those $u = 1-\theta$ vanishes identically, so the coefficient bijection is exactly the uniqueness of the unbiased estimator of each; that is the only step the richness hypothesis carries, and Remark A.16 exhibits a class where uniqueness fails without it. Solving for the two named functionals: $h(u) = 1-u$ forces $t(1,1) = 1$ and all other values 0, i.e. $t = \mathbf{1}\{\tau = 1,\ X_\tau = 1\} = X_1$; $h(u) = 1-u^k$ (which requires $N_{\max} \ge k$) forces $t(j,1) = \mathbf{1}\{j \le k\}$ and $t(N_{\max},0) = 0$, i.e. $t = \mathbf{1}\{\tau \le k,\ X_\tau = 1\}$. Both computations are carried out at the level of h, so unbiasedness holds for every mixing law. That $\mathbf{1}\{\tau = 1,\ X_\tau = 1\}$ is the unique unbiased estimator of a success probability under inverse binomial sampling is classical; what is used here is the spanning identity, which says the truncation discards nothing.

Proposition A.17. Work in the general class, where the joint law of $X_1,\dots,X_{N_{\max}}$ is unrestricted. By hypothesis there is a terminal history $h = (0,\dots,0)$ of length $m < k$, reached with $\mathbb{P}(h) > 0$, on which the rule stops. Because τ is a function of the history alone and carries no randomisation of its own, the law of H_τ is the law of the realised prefix, and no realised prefix extends h. The conditional law of $(X_{m+1},\dots,X_{N_{\max}})$ given h is therefore unconstrained by the law of H_τ: let P_0 set $X_{m+1} = \cdots = X_{N_{\max}} = 0$ almost surely on h and let P_1 set $X_{m+1} = 1$ almost surely on h, leaving both laws unchanged elsewhere. The two induce the same law of H_τ, both lie in the class, and their values of $\mathrm{succ}@k$ differ by at least $\mathbb{P}(h) > 0$ because $m+1 \le k$. Hence $\mathrm{succ}@k$ of the potential sequence is not identified. The minimal instance is $\tau \equiv 1$, which is deterministic and already inside the latent-variable class of Definition 3.1: there $H_\tau = (X_1)$, so the observation law depends on π only through μ_1, while $\mathrm{succ}@k = 1 - \mathbb{E}_\pi[(1-\theta)^k]$ is not a function of μ_1 alone for $k \ge 2$ — two mixing laws with equal means and different $\mathbb{E}[(1-\theta)^k]$ exist. Both restrictions in the statement are needed, and Remark A.18 gives the two counterexamples that show it: a randomised rule with positive continuation probability everywhere admits a Horvitz–Thompson estimator, and inside the exchangeable class the terminal histories of a first-failure rule already span $\mathrm{succ}@2$; its third item shows that the observation cannot be compressed to (τ, X_τ) once the rule is not the first-success rule. The identified quantity is immediate: $\mathbf{1}\{\exists\, i \le \tau \wedge k : X_i = 1\}$ is a function of H_τ, and its expectation is $\mathrm{succ}@k$ of the truncated procedure by Proposition 3.4 applied to that procedure; when $\mathbb{P}(\tau < k, X_1 = \cdots = X_\tau = 0) = 0$ the indicator equals $\mathbf{1}\{X_1 \vee \cdots \vee X_k = 1\}$ almost surely and the two quantities coincide. The operational reading is that the estimand must be indexed by the procedure that is actually run, stopping rule included.

Theorem 3.5. By (ii) and Lemma A.3 with $n = k$, the orbit averages of T satisfy $B_c(p) = 1-(1-p)^k$ on an infinite set, hence identically, so the rigidity computation of Corollary A.12 at $n = k$ gives $c_s = 1 - \binom{k-s}{k}/\binom{k}{k} = \mathbf{1}\{s \ge 1\}$. The orbit $\{|x| = 0\}$ is a single point, so $T(0,\dots,0) = c_0 = 0$. On each orbit with $s \ge 1$ the average of T is 1. Under (iii) we have $T \le 1$ pointwise, and a function bounded by 1 with average 1 on a finite set equals 1 there. Under (iii′) the orbit $\{|x| = k\}$ is a single point, so $T(1,\dots,1) = c_k = 1$, and coordinatewise monotonicity gives $T \le T(1,\dots,1) = 1$ everywhere, after which the same averaging argument applies. Either way $T = \mathbf{1}\{x \ne 0\}$ and $F(G) = \mathbb{P}_G(\exists i \le k : X_i = 1) = \mathrm{succ}@k(G)$.

Dropping (iii) and (iii′) together breaks uniqueness, and boundedness is not a substitute for the range restriction, every function on the finite set $\{0,1\}^k$ being bounded. For $k = 2$ and $a \in \mathbb{R}$ set $T_a(0,0) = 0$, $T_a(1,1) = 1$, $T_a(1,0) = 1+a$, $T_a(0,1) = 1-a$. Under any i.i.d. law the two mixed outcomes are equally likely, so the a-terms cancel and T_a satisfies (i) and (ii) for every a; but on a non-exchangeable G, $F_a(G) = \mathrm{succ}@k(G) + a\big[\mathbb{P}_G(1,0) - \mathbb{P}_G(0,1)\big]$, which differs from $\mathrm{succ}@k$. These antisymmetric distortions are excluded by (iii) and by (iii′) alike, since $a \ne 0$ makes T_a leave $[0,1]$ on one orbit point and decrease in one coordinate. Uniqueness also fails when more than k attempts are observed: for $n > k$ the orbit averages $c_s = 1 - \binom{n-s}{k}/\binom{n}{k}$ are 0 at $s = 0$ and 1 for $s > n-k$, but lie strictly

inside $(0, 1)$ for $1 \le s \le n-k$, where they no longer pin T down pointwise; this is why the statement is tied to budget-k observation. □

A.4 Proofs for Section 5

This subsection proves the results of Section 5 on the reporting protocol, the variance floor and the allocation, as stated in §A.1.

Proof of Proposition A.22. (a) Condition on $\theta = (\theta_1, \dots, \theta_R)$. The runs are conditionally independent with $S_r \sim \mathrm{Bin}(n, \theta_r)$, so $S_r/n \to \theta_r$ almost surely. Write $M := nR$ for the pooled number of attempts and $D := \sum_r S_r$ for the pooled number of successes, so that $D/M \to \bar\theta := R^{-1}\sum_r \theta_r$. For fixed k,

$$\frac{\binom{M-D}{k}}{\binom{M}{k}} \;=\; \prod_{i=0}^{k-1} \frac{M-D-i}{M-i} \;=\; \prod_{i=0}^{k-1} \frac{1-D/M-i/M}{1-i/M} \;\longrightarrow\; (1-\bar\theta)^k,$$

so $\widehat{C}_{\mathrm{pooled}} \to L := 1-(1-\bar\theta)^k$ almost surely, hence in probability after integrating over θ. Because R is fixed and $\mathrm{Var}(\theta) > 0$, $\bar\theta$ is non-degenerate, and $t \mapsto 1-(1-t)^k$ is strictly increasing on $[0,1]$, so $\mathrm{Var}(L) > 0$. Pooling erases the run labels, and in the limit the pooled estimator behaves like the plug-in evaluated at the pooled latent rate.

(b) Fix $\varepsilon > 0$. On the event $\{\mathrm{width}(I_n) \le \varepsilon\} \cap \{c \in I_n\}$ the interval contains both c and $\widehat{C}_{\mathrm{pooled}}$, so $|\widehat{C}_{\mathrm{pooled}} - c| \le \varepsilon$; hence

$$\{c \in I_n\} \;\subseteq\; \{|\widehat{C}_{\mathrm{pooled}} - c| \le \varepsilon\} \,\cup\, \{\mathrm{width}(I_n) > \varepsilon\}.$$

The second event has probability tending to 0, and $\{x : |x-c| \le \varepsilon\}$ is closed, so the portmanteau theorem applied to $\widehat{C}_{\mathrm{pooled}} \Rightarrow L$ gives $\limsup_n \mathbb{P}(c \in I_n) \le \mathbb{P}(|L-c| \le \varepsilon)$. Letting $\varepsilon \downarrow 0$ and using continuity from above of the probability measure along the decreasing events $\{|L-c| \le \varepsilon\}$, whose intersection is $\{L = c\}$, yields $\limsup_n \mathbb{P}(c \in I_n) \le \mathbb{P}(L = c)$, and $\mathbb{P}(L = c) \le \max_x \mathbb{P}(L = x) < 1$ because L is non-degenerate. When c is not an atom of L — in particular whenever π has a density, so that L has none — the bound is 0.

The route through $\mathbf{1}\{c \in I_n\} \to \mathbf{1}\{L = c\}$ is not available: the indicator of a closed set is discontinuous exactly on its boundary, and $\{L = c\}$ is that boundary. This is why the conclusion is a lim sup bound that can be strictly positive. A nonzero atom permits a positive upper bound; whether a particular interval attains it needs separate analysis. A two-point mixing law with an even number of runs is the canonical case: $\bar\theta$ can equal μ_1 exactly, which places an atom of L precisely at $\mathrm{pass@}k^{\mathrm{iid}}$. □

Proof of Proposition A.21 and Corollary A.23. Let S be a uniformly random k-subset of the M pooled positions, drawn independently of everything else. Applied to the pooled attempt vector, the subset identity in the proof of Proposition 3.3 gives $\widehat{C}_{\mathrm{pooled}} = \mathbb{P}(\exists\,(r,i) \in S : X_i^{(r)} = 1 \mid X)$, so that $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] = 1 - \mathbb{P}(\text{every position of } S \text{ fails})$. Write $c_r := \#\{i : (r,i) \in S\}$; because S is uniform, c is multivariate hypergeometric with the weights stated in (10). Conditionally on c and on $\theta = (\theta_1, \dots, \theta_R)$ the chosen attempts are independent across runs and conditionally i.i.d. $\mathrm{Bern}(\theta_r)$ inside run r, so all of them fail with probability $\prod_r (1-\theta_r)^{c_r}$. The θ_r are i.i.d. π, so averaging over them gives $\prod_r m_{c_r}$, and averaging over c gives (10).

For the corollary, $u \mapsto u^c$ is convex on $[0,1]$, so Jensen's inequality gives $m_c \ge m_1^c$, strictly unless $c \le 1$ or θ is almost surely constant. Hence $\prod_r m_{c_r} \ge \prod_r m_1^{c_r} = (1-\mu_1)^k$ for every allocation, with equality exactly when every $c_r \le 1$ or $\mathrm{Var}(\theta) = 0$; taking the expectation over c and subtracting from 1 gives $\mathbb{E}[\widehat{C}_{\mathrm{pooled}}] \le 1-(1-\mu_1)^k = \mathrm{pass@}k^{\mathrm{iid}}$. Equality throughout requires either $\mathrm{Var}(\theta) = 0$ or that every allocation of positive probability be collision-free, and the latter happens exactly when $k = 1$ or $n = 1$; in the case $n = 1$ the pooled estimator is defined only for $M = R \ge k$, which recovers the sufficiency direction of Theorem 4.1. □

Proof of Theorem A.20. General Z. Given Z the continuations are exchangeable, so Proposition 3.3 applies conditionally and $\mathbb{E}[\widehat{C}_{n,k} \mid Z] = \mathbb{P}(X_1 \vee \cdots \vee X_k = 1 \mid Z)$; averaging over Z gives unbiasedness of $\widehat{C}_{\mathrm{run}}$ for $\mathrm{succ@}k$. The complete run vectors $(Z_r, X^{(r)})$ being i.i.d., the law of total vari-

ance applied to the average of the R independent copies $\widehat{C}_{n,k}(X^{(r)})$, conditioning on $(Z_1, \dots, Z_R)$, gives (9); this is the argument below with $g(\theta_r)$ replaced by $\mathbb{P}(X_1 \vee \cdots \vee X_k = 1 \mid Z_r)$.

Definition 3.1. Condition on $\theta = (\theta_1, \dots, \theta_R)$. Given θ_r the attempts of run r are i.i.d. $\mathrm{Bern}(\theta_r)$, an exchangeable law, so Proposition 3.3 gives $\mathbb{E}[\widehat{C}_{n,k}(X^{(r)}) \mid \theta_r] = 1 - (1-\theta_r)^k = g(\theta_r)$. Hence $\mathbb{E}[\widehat{C}_{\mathrm{run}} \mid \theta] = R^{-1} \sum_r g(\theta_r)$, whose variance is V_0/R because the θ_r are i.i.d.; and, the runs being conditionally independent, $\mathbb{E}\big[\mathrm{Var}(\widehat{C}_{\mathrm{run}} \mid \theta)\big] = R^{-2} \sum_r \mathbb{E}\big[\mathrm{Var}(\widehat{C}_{n,k} \mid \theta_r)\big] = W_\pi(n,k)/R$. The law of total variance gives (9). V_0 is a functional of π and k alone, so it does not involve n.

Monotonicity: for fixed θ, $\widehat{C}_{n,k}$ is a U-statistic of fixed order k computed from n i.i.d. draws, and the variance of a U-statistic is non-increasing in the sample size (Hoeffding, 1948); averaging over θ preserves this. Convergence to zero: $S/n \to \theta$ almost surely gives $\widehat{C}_{n,k} \to g(\theta)$ almost surely, and $\widehat{C}_{n,k} \in [0,1]$, so bounded convergence gives $\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) \to 0$ and then $W_\pi(n,k) \to 0$.

Lower bound: under the conditions of Proposition A.14, part (iii) of that proposition makes $\widehat{C}_{\mathrm{run}}$ the UMVUE of succ@k, so any unbiased T has $\mathrm{Var}_\pi(T) \geq \mathrm{Var}_\pi(\widehat{C}_{\mathrm{run}}) = (V_0 + W_\pi(n,k))/R \geq V_0/R$. □

Proof of Proposition A.25. Given θ the attempts are i.i.d., so the variance decomposition for a U-statistic of order k applies (the theory of Hoeffding, 1948): $\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\zeta_c(\theta)$ with $\zeta_c = \mathrm{Var}\big(\mathbb{E}[h \mid X_1, \dots, X_c]\big)$. For the kernel $h(x) = 1 - \prod_{i \leq k}(1 - x_i)$,

$$\mathbb{E}[h \mid X_1, \dots, X_c] = 1 - (1-\theta)^{k-c}\, \mathbf{1}\{X_1 = \cdots = X_c = 0\},$$

and $\mathbf{1}\{X_1 = \cdots = X_c = 0\} \sim \mathrm{Bern}\big((1-\theta)^c\big)$, so $\zeta_c(\theta) = (1-\theta)^{2(k-c)}(1-\theta)^c\big(1 - (1-\theta)^c\big)$, which is (13). Only the $c = 1$ term survives the scaling by n, since $\binom{k}{1}\binom{n-k}{k-1}/\binom{n}{k} = k^2/n + O(n^{-2})$ while the terms with $c \geq 2$ are $O(n^{-c})$; hence $n\,\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) \to k^2\zeta_1(\theta)$. Equivalently, since $\widehat{C}_{n,k}$ depends on the data only through $S \sim \mathrm{Bin}(n,\theta)$, the same quantity is $\sum_s \binom{n}{s}\theta^s(1-\theta)^{n-s}\widehat{C}_{n,k}(s)^2 - g(\theta)^2$; the two expressions agree identically. □

Proof of Lemma A.26. Fix θ, write $u := 1 - \theta$ and $Z_i := (1 - X_i) - u$, so that $Z_1, \dots, Z_n$ are i.i.d., centred, with $\mathbb{E}[Z_i^2] = \theta u$. Since $\widehat{C}_{n,k} = 1 - \binom{n}{k}^{-1} \sum_{|S|=k} \prod_{i \in S}(1 - X_i)$, expanding each product as $\prod_{i \in S}(u + Z_i) = \sum_{T \subseteq S} u^{k-|T|} Z_T$ with $Z_T := \prod_{i \in T} Z_i$ and collecting the k-subsets containing a given T with $|T| = j$ — there are $\binom{n-j}{k-j}$ of them, and $\binom{n-j}{k-j}/\binom{n}{k} = \binom{k}{j}/\binom{n}{j}$ — gives

$$\widehat{C}_{n,k} \;=\; \mathrm{const} \;-\; \sum_{j=1}^{k} \frac{\binom{k}{j}}{\binom{n}{j}}\, u^{k-j} \sum_{|T|=j} Z_T.$$

Distinct Z_T are orthogonal and $\mathbb{E}[Z_T^2] = (\theta u)^j$, and there are $\binom{n}{j}$ sets of size j, so

$$\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) \;=\; \sum_{j=1}^{k} \frac{\binom{k}{j}^2}{\binom{n}{j}^2}\, u^{2(k-j)} (\theta u)^j \binom{n}{j} \;=\; \sum_{j=1}^{k} \frac{\binom{k}{j}^2}{\binom{n}{j}}\, \theta^j (1-\theta)^{2k-j},$$

and averaging over π gives (14). This is the orthogonal form of the same Hoeffding decomposition whose first form is (13); the two agree term by term in n for each k.

For the shape, $1/\binom{n}{j} = j! \Big/ \prod_{\ell=0}^{j-1}(n - \ell)$ and $n/\binom{n}{j} = j! \Big/ \prod_{\ell=1}^{j-1}(n - \ell)$, both of the form $q(n) = c \prod_{\ell \in L}(n - \ell)^{-1}$ with L a finite set of integers strictly below n and $c > 0$. Any such q satisfies

$$q'(n) = -q(n) \sum_{\ell \in L} \frac{1}{n - \ell} \;<\; 0, \qquad q''(n) = q(n)\Big[\Big(\sum_{\ell \in L} \frac{1}{n-\ell}\Big)^2 + \sum_{\ell \in L} \frac{1}{(n-\ell)^2}\Big] \;>\; 0,$$

by differentiating $\log q$ twice, with $q \equiv c$ and $q' = q'' = 0$ when $L = \emptyset$, which is the case $j = 1$ of $n/\binom{n}{j}$. Since $a_j \geq 0$, both $W_\pi(\cdot, k)$ and $n \mapsto nW_\pi(n,k)$ are non-negative combinations of such q and therefore non-increasing and convex; $(\kappa+n)V_0$ is affine and $(\kappa+n)W_\pi(n,k) = nW_\pi(n,k) + \kappa W_\pi(n,k)$ is a sum of two convex functions, so Φ is convex. The affine term is increasing whenever $V_0 > 0$,

which is why convexity and not monotonicity is what Φ inherits; the stated counterexample is the case $W_\pi \equiv 0$, where Φ is exactly that affine term. For strictness, $L \neq \emptyset$ in $n/\binom{n}{j}$ exactly when $j \geq 2$, and $a_j = \binom{k}{j}^2 \mathbb{E}_\pi[\theta^j(1-\theta)^{2k-j}] > 0$ for such a j exactly when π charges $(0,1)$, which needs $k \geq 2$ for any $j \geq 2$ to exist.

Finally, a convex function restricted to the integers has non-decreasing increments, so $\min_{n \geq k} \Phi(n)$ over integers is attained at $n = k$ if and only if $\Phi(k+1) - \Phi(k) \geq 0$, which is (15). □

Proof of Proposition A.24. A design of cost N has $R = N/(c_0 + nc_1)$, so $\mathrm{Var}(\widehat{C}_{\mathrm{run}}) = (V_0 + W_\pi(n,k))(c_0 + nc_1)/N = (c_1/N)\,\Phi(n)$, and minimising the variance at fixed cost is minimising Φ. If $V_0 > 0$ then $\Phi(n) \geq V_0(\kappa + n) \to \infty$ while $\Phi(k) < \infty$, so only finitely many n can beat $\Phi(k)$ and the minimum is attained at a finite $n^\star \geq k$.

(i) As $\kappa \to 0$ the objective is $f(n) := n\big(V_0 + W_\pi(n,k)\big)$. If $n^\star = k$ then in particular $f(k) \leq f(k+1)$, which rearranges to $V_0 \geq kW_\pi(k,k) - (k+1)W_\pi(k+1,k)$. Conversely, $f(n) = nV_0 + nW_\pi(n,k)$ is a sum of a linear term and $n \mapsto nW_\pi(n,k)$, which is convex by Lemma A.26; so f is convex, its increments are non-decreasing, and $f(k) \leq f(k+1)$ forces $f(k) \leq f(n)$ for all $n > k$, i.e. $n^\star = k$. The same argument runs at any $\kappa \geq 0$ on Φ itself, which the lemma also makes convex, and gives the boundary test (15). At $k = 2$ the convexity is visible in closed form, $n\,\mathrm{Var}(\widehat{C}_{n,k} \mid \theta) = 4\zeta_1(\theta) + 2\theta^2(1-\theta)^2/(n-1)$ being convex in n for each θ and therefore after mixing; and the exact-rational sweep on 505 mixing laws and 83,325 grid points for $k \in \{8, 16, 32\}$, where the second difference is never negative (Table S1; a double-precision sweep of the same grid returns 2,154 spurious negatives, all of them below the resolution of the closed form), agrees with the lemma on every cell it covers.

(ii) Assume $\pi((0,1)) > 0$. At $n = k$ the classical estimator is the indicator $\mathbf{1}\{S \geq 1\}$, so $W_\pi(k,k) = \mathbb{E}_\pi\big[(1-\theta)^k\big(1-(1-\theta)^k\big)\big]$, which is strictly positive exactly when π charges $(0,1)$; that is the stated equivalence. Strict monotonicity of $W_\pi(\cdot,k)$: the estimator on $n+1$ attempts is the average of the $n+1$ leave-one-out copies of the estimator on n attempts, because a given k-subset of $[n+1]$ lies in exactly $n+1-k$ of the $n+1$ leave-one-out samples and $\binom{n+1}{k}^{-1} = (n+1-k)\big[(n+1)\binom{n}{k}\big]^{-1}$. Given θ those copies are identically distributed, so the conditional variance is non-increasing in n, and equality would force them to coincide almost surely; that fails at every $\theta \in (0,1)$, since on the positive-probability event that exactly one of the $n+1$ attempts succeeds, deleting the successful attempt leaves the estimator at 0 while deleting any other leaves it at $k/n > 0$. Averaging over π preserves the strict decrease because $\pi((0,1)) > 0$. Now for every n,

$$\Phi(n) - \Phi(n+1) \;=\; \kappa\big[W_\pi(n,k) - W_\pi(n+1,k)\big] \;-\; \big[(n+1)\big(V_0 + W_\pi(n+1,k)\big) - n\big(V_0 + W_\pi(n,k)\big)\big].$$

Fix $n_0 \geq k$. The subtracted bracket is bounded over $k \leq n \leq n_0$ and does not involve κ, while $\min_{k \leq n \leq n_0}\big[W_\pi(n,k) - W_\pi(n+1,k)\big]$ is strictly positive, so for κ large enough $\Phi(n) > \Phi(n+1)$ for every $n \leq n_0$, whence $n^\star > n_0$; letting n_0 grow gives $n^\star \to \infty$. If instead π is supported on $\{0,1\}$, every run is all successes or all failures, the classical estimator is a deterministic function of θ, $W_\pi \equiv 0$, and $\Phi(n) = V_0(\kappa + n)$ with $V_0 = \pi(\{1\})\big(1 - \pi(\{1\})\big) > 0$ is strictly increasing in n, so $n^\star = k$ at every κ: the divergence claim needs the hypothesis. At fixed R, by (9) and $W_\pi(n,k) \downarrow 0$, $\inf_n \mathrm{Var}(\widehat{C}_{\mathrm{run}}) = V_0/R$, which no choice of c_0 and c_1 changes.

(iii) $V_0 = \mathrm{Var}_\pi(g(\theta))$ vanishes exactly when $g(\theta)$ is almost surely constant, and g is strictly increasing on $[0,1]$, so $V_0 = 0$ if and only if $\pi = \delta_{\theta_0}$ for some θ_0 — that is, only when the shared intermediate object is deterministic and there is no within-run dependence. Then $\Phi(n) = W_\pi(n,k)(\kappa + n) = \kappa\,W(n) + n\,W(n)$, and at $\theta_0 \in (0,1)$ Lemma A.26 makes both terms non-increasing. The first is strictly decreasing when $\kappa > 0$, by the strict monotonicity of $W_\pi(\cdot,k)$ proved in (ii), whose hypothesis $\pi((0,1)) > 0$ holds here; the second is strictly decreasing when $k \geq 2$, by the strictness clause of the lemma. So under either condition Φ is strictly decreasing and the optimum puts the entire budget inside one run. Neither holds at $k = 1$ with $\kappa = 0$, and there the conclusion is false rather than merely unproved: $W_\pi(n,1) = \theta_0(1-\theta_0)/n$ by (14), so $\Phi(n) = \theta_0(1-\theta_0)$ for every n and the relaxation is indifferent. At $\theta_0 \in \{0,1\}$ every run is all failures or all successes, so $W \equiv 0$ and $\Phi \equiv 0$ and no n is preferred; that is the degenerate case named in the statement. □

The allocation rule. This is a two-stage sampling problem, not a new principle of experimental design, and allocating a budget across questions is prior work (Kazdan et al., 2025, Thm. 1); the axis

Table 5: The cost-optimal number of continuations per fresh run at budget $k = 4$ grows with the cost ratio and depends on the shape of the mixing law, not only on its first two moments. Two moment pairs $(\mu_1, \mathrm{Var}(\theta))$, one high-rate and one low-rate, each realised twice: as the two-point mixture carrying those two moments with one atom at the boundary of $[0, 1]$ nearer the mean, and as the Beta distribution carrying them. The pairs are those of the headline shared-prefix design and of the reasoning tree, with $\mathrm{Var}(\theta) = \hat{\rho}\,\mu_1(1-\mu_1)$ at the median $\hat{\rho}$ over the questions on which it is defined (Table 7); supplementary Table S9 matches the same corpora at the mean $\hat{\rho}$ instead. Columns 2–3 give the floor V_0 of (5) and the within-run term at $n = k$; the next four give $n^\star$ of Proposition A.24 at four cost ratios $\kappa = c_0/c_1$; the last four give $\mathrm{Var}(\widehat{C}_{\mathrm{run}})$ at $n^\star$ in units of the floor V_0/R, so that 1.00 would mean the design has bought everything the floor allows. The floor is sensitive to the shape of the mixing law and not only to its first two moments: the two-point rows are the extreme case, and the Beta rows are the ones to read. All entries are exact computations from (13); no simulation and no experimental input enter this table.

			$n^\star$ at $\kappa = c_0/c_1$				$\mathrm{Var}(\widehat{C}_{\mathrm{run}})$ at $n^\star$, in units of V_0/R			
Mixing law	V_0	$W(k)$	0.1	1	10	100	0.1	1	10	100
two-point (0.90, 0.020)	1.5×10^{-5}	2.7×10^{-3}	28	30	40	88	3.87	3.62	2.81	1.72
two-point (0.38, 0.068)	0.201	2.5×10^{-2}	4	4	4	7	1.12	1.12	1.12	1.03
Beta (0.90, 0.020)	1.1×10^{-3}	5.5×10^{-3}	7	8	12	30	2.66	2.35	1.77	1.26
Beta (0.38, 0.068)	0.090	0.119	4	4	8	20	2.32	2.32	1.51	1.18

priced here is the one inside a question, between fresh executions and attempts that share a latent state. Attempts buy W; only fresh runs buy V_0, so the design $R = 1$ never falls below V_0. At fixed cost the optimum minimises $\Phi(n) := (V_0 + W_\pi(n, k))(\kappa + n)$ over $n \geq k$; Proposition A.24 gives a closed-form criterion for $n^\star = k$, on a convexity of $W_\pi(\cdot, k)$ that Lemma A.26 proves at every k, and at the measured κ the relaxed optimum is 9 or 5 continuations on the two fitted laws at the point estimate of κ (Table 37b); these are conditional allocations, not a precise cost optimum over the uncertainty in κ.

Table 6: At a fixed total cost the integer optimum can differ from the relaxed $n^\star$ of Proposition A.24 in either direction, and it depends on the total cost where the relaxation does not. The allocation on the integer feasible region rather than on the continuous relaxation it is stated for, at $k = 4$. A design of total cost N attempts can only take $R = \lfloor N/(\kappa+n) \rfloor$ runs, and that floor is what a practitioner faces: the relaxation's $n^\star$ minimises $\Phi(n) = (V_0+W_\pi(n,k))(\kappa+n)$ and does not depend on N at all, while the integer optimum minimises $(V_0+W_\pi(n,k))/R$ over the same laws and does depend on N. The four mixing laws, and the closed form for V_0 and W_π, are those of Table 5, whose cells this computation reproduces. Panel (a) sweeps the total cost at the measured cost ratio and panel (b) sweeps the cost ratio at a fixed total cost. Where the two disagree the integer answer is the one a fixed total cost can actually buy.

(a) total cost swept at the measured cost ratio $\kappa = 2.96$

Total cost N	two-point (0.90, 0.020)	two-point (0.38, 0.068)	Beta (0.90, 0.020)	Beta (0.38, 0.068)
16	(1, 13)	(2, 5)	(1, 13)	(2, 5)
32	(1, 29)	(4, 5)	(2, 13)	(4, 5)
64	(2, 29)	(9, 4)	(4, 13)	(8, 5)
128	(4, 29)	(18, 4)	(9, 11)	(16, 5)
256	(8, 29)	(36, 4)	(21, 9)	(32, 5)
512	(16, 29)	(73, 4)	(39, 10)	(64, 5)
1,024	(31, 30)	(147, 4)	(79, 10)	(128, 5)
continuous $n^\star$	32	4	9	5

(b) cost ratio swept at total cost $N = 128$ attempts; each pair is the integer optimum and the relaxation's $n^\star_{\mathrm{rel}}$

	two-point (0.90, 0.020)		two-point (0.38, 0.068)		Beta (0.90, 0.020)		Beta (0.38, 0.068)	
κ	$(R^\star, n^\star)$	$n^\star_{\mathrm{rel}}$	$(R^\star, n^\star)$	$n^\star_{\mathrm{rel}}$	$(R^\star, n^\star)$	$n^\star_{\mathrm{rel}}$	$(R^\star, n^\star)$	$n^\star_{\mathrm{rel}}$
0.5	(5, 25)	29	(28, 4)	4	(17, 7)	7	(28, 4)	4
2.96	(4, 29)	32	(18, 4)	4	(9, 11)	9	(16, 5)	5
10	(2, 54)	40	(9, 4)	4	(6, 11)	12	(7, 8)	8
50	(1, 78)	67	(2, 14)	6	(2, 14)	22	(2, 14)	15

The integer feasible region. The relaxation treats R as continuous. Table 6 enumerates the integer region $R = \lfloor N/(\kappa + n) \rfloor \geq 1$, $n \geq k$, at the four laws of Table 5: at $N = 128$ and the measured $\kappa = 2.96$ the integer optimum takes $n^\star = 29$, 4, 11 and 5 continuations where the relaxation gives 32, 4, 9 and 5; at $N = 16$ the first law's integer optimum drops to 13, and the sweep over κ at a fixed total cost moves the two apart in both directions. The relaxation is a large-budget statement; the table is the one a fixed total cost can buy.

B Complete evidence

This appendix holds, mechanism by mechanism, the source tables for the numbers of Section 6. Table 7 gives one row per design and Figure 3 sets each measured gap against the second-order reference of (6); one subsection per mechanism follows. Robustness checks that no main-text sentence rests on, the full grids and the figures that repeat these tables are in the supplementary report (Supplement §S2).

B.1 Reading the evidence summary

This subsection says how to read the rows of Table 7. The shared-prefix 7B row is the headline design and the other seven corroborate it, so the eight rows are not eight independent tests. A correlation is undefined where a question's runs agree — on 138 of the 200 headline-design questions and 141 of the 194 tree-search ones, which contribute a measured gap of exactly zero and stay in every denominator; the tree row's gap is its $R_q \geq 4$ subset, 4,880 of 8,061 questions. Every block measures the gap the same way, by the $\widehat{U}_k - \overline{Y}$ of Proposition A.8 and the columns the caption defines; Appendix C holds the controlled-truth simulation that checks these predictions against a known truth, and the supplementary report the full grids (Supplement §§S2–S4).

Table 7: At $k = 4$ the estimand gap has a positive point estimate in all eight designs and an interval above zero in seven (all but the self-debugging 7B loop), and run-level runCS contains the held-out reference more often than the attempt-level bootstrap in all eight. One row per design — the eight designs of §6 (full grids in Supplement §§S2–S4; Appendix B.1 explains how to read the rows). Columns: corpus and model; design as independent runs × attempts; median within-run intraclass correlation where defined; $\hat{\mu}_1$; the estimand gap $\widehat{U}_k - \overline{Y}$ (Proposition A.8, unbiased for pass@k^{iid} − succ@k) at $k = 4$ with a 95% question-bootstrap interval (bootstrap seed differs; canonical interval in Table 50) and at the largest budget; the ratio of the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ to its finite-design prediction, $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ of Proposition A.21 less $\widehat{C}_{\text{run}}$, ranged over $k \geq 2$ (the two sides subtract different estimates of succ@k); and three containment figures at $k = 4$: an attempt-level bootstrap and run-level runCS (the Beta-mixture confidence sequence on run indicators, (17)), both against a held-out $\overline{Y}_4$, and the first-attempt sequence carried through g_4, against the held-out cross-run statistic for pass@k^{iid}: $\widehat{U}_k$ on the four shared-prefix rows and its first-attempt form on the reasoning-tree, self-debugging and tree-search rows. The gap columns use the first-attempt form on the self-debugging and tree-search rows and the random-sibling form on the reasoning tree (§6.2). All references carry noise, so these are containment figures, not coverage of a known parameter. A dash marks a quantity the design does not define: $\hat{\rho}$ is not reported for the self-debugging loops, whose attempts are ordered, and the finite-design ratio is reported only for the four shared-prefix arms, whose attempts are exchangeable.

Design and estimand gap

Corpus / model	$R \times n$ design	$\hat{\rho}$ median	$\hat{\mu}_1$	$\widehat{U}_4 - \overline{Y}$ 95% CI
Shared prefix, 7B (headline) *GSM8K, generated*	16×8	0.221	0.8959	+0.0262 [+0.0156, +0.0375]
Shared prefix, 1.5B, Arm A *GSM8K, off saturation*	32×32	0.663	0.5368	+0.1678 [+0.1480, +0.1865]
Shared prefix, 7B, Arm B *competition maths*	32×32	0.075	0.7013	+0.0360 [+0.0174, +0.0581]
Shared prefix, 1.5B, Arm C *no admissibility condition*	32×32	0.750	0.5341	+0.1814 [+0.1628, +0.2008]
Reasoning tree *public, human labels*	$R_q \geq 4 \times 4$	0.289	0.3834	+0.1668 [+0.1628, +0.1707]
Self-debugging loop, 7B *HumanEval, generated*	$16\times \leq 4$	—	0.8190	+0.0092 [−0.0116, +0.0271]
Self-debugging loop, 1.5B *HumanEval, generated*	$16\times \leq 4$	—	0.5271	+0.1253 [+0.0964, +0.1526]
Tree search, 7B *GSM8K, generated*	16×4	0.205	0.8934	+0.0134 [+0.0016, +0.0257]

Intervals in every evidence table are per cell, resampled over questions, with no multiplicity adjustment. They are redrawn for each table with its own random seeds, so the same statistic can carry slightly different interval endpoints in different tables; the point estimates agree, and Table 50 holds the canonical interval.

B.2 Continuations of a shared reasoning prefix

This section is the complete version of Section 6.1 of the main text.

Design. For each question we sample $R = 16$ prefixes of $P = 48$ tokens and, from each prefix, $n = 8$ independent continuations, so that a prefix is a run and a continuation is an attempt. The instruction, both few-shot exemplars, the answer-extraction and normalisation rules and the chat-template construction are reproduced verbatim in `repro/prompt.md` of the supplementary material; the sampling parameters are temperature 0.7, top-p 1.0, top-k disabled, repetition penalty 1.0, and a limit of 512 new tokens per attempt, of which the shared prefix is the first 48. A prefix is admissible only if it does not already contain the final answer and does not terminate before P tokens. The reference for the containment experiment is a held-out half of the prefixes.

Table 7: Evidence summary (continued): largest budget, pooling check and the three interval comparisons.

Largest budget, pooling check, and the three interval comparisons

Corpus / model	$\widehat{U}_k - \overline{Y}$ at $k = k_{\max}$	Prop. A.21 ratio range	succ@k, $k=4$ attempt / RUNCS	pass@k^{iid}, $k=4$ $g_4(\text{RUNCS}_1)$
Shared prefix, 7B (headline) *GSM8K, generated*	+0.0266 ($k=8$)	0.87–1.12	0.864 / 0.966	0.954
Shared prefix, 1.5B, Arm A *GSM8K, off saturation*	+0.2410 ($k=32$)	1.01–1.04	0.146 / 0.892	0.868
Shared prefix, 7B, Arm B *competition maths*	+0.0287 ($k=32$)	0.95–1.01	0.208 / 0.922	0.965
Shared prefix, 1.5B, Arm C *no admissibility condition*	+0.2596 ($k=32$)	1.02–1.03	0.190 / 0.905	0.878
Reasoning tree *public, human labels*	+0.1668 ($k=4$)	—	0.609 / 0.788	0.428
Self-debugging loop, 7B *HumanEval, generated*	+0.0092 ($k=4$)	—	0.785 / 0.925	0.824
Self-debugging loop, 1.5B *HumanEval, generated*	+0.1253 ($k=4$)	—	0.670 / 0.887	0.728
Tree search, 7B *GSM8K, generated*	+0.0134 ($k=4$)	—	0.913 / 0.984	0.900

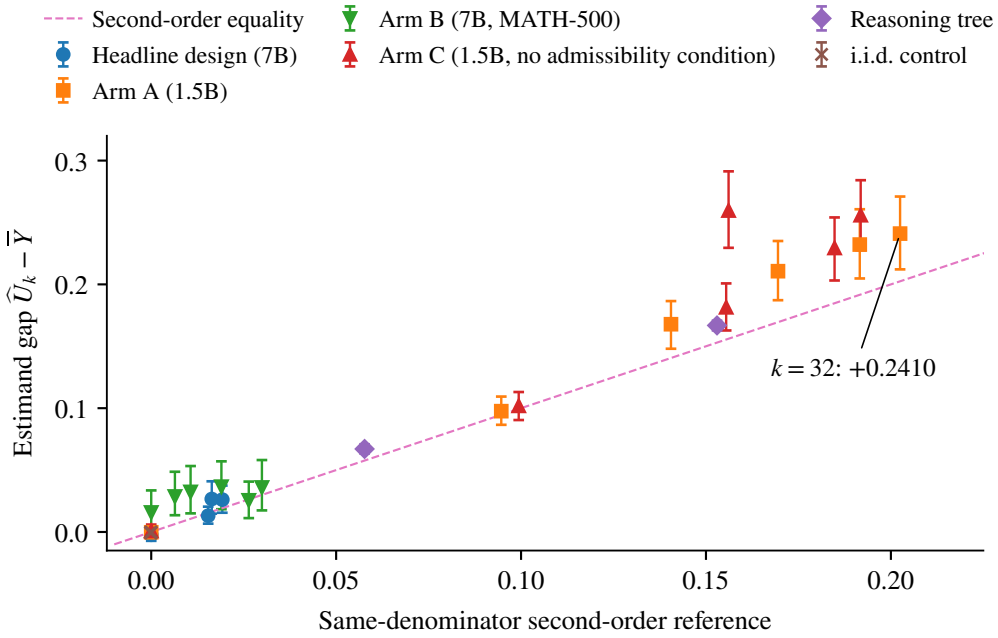


Figure 3: The second-order expression is a reference for the estimand gap, not a controlled approximation in the measured regimes. One point per corpus and budget shows the reported estimand gap with its 95% question-bootstrap interval for $\widehat{U}_k - \overline{Y}$ against the same-denominator reference of (6), that is, the second-order reference averaged over the same questions as the gap. The i.i.d. control remains near the origin; Fig. S2 enlarges those points without changing their coordinates.

The within-run correlation is real and widely spread: across the 62 of the 200 questions on which it is defined, the intraclass correlation has median 0.22 with quartiles 0.02 and 0.65, so some questions lock onto their prefix almost completely and others not at all; on the other 138 every continuation

Table 8: On the headline design the pooled excess is positive from $k = 2$ on, with a question-bootstrap interval above zero, while the $k = 1$ control contains zero. Within-run dependence and the pooled excess (16 prefixes × 8 continuations per question on 200 mathematical reasoning questions, exact answer match as verifier). Panel (a) reports the distribution of the within-run correlation $\hat{\rho}$ on two denominators: across the 62 questions on which it is defined, and across all 200 with the 138 undefined entries set to zero as an explicitly imputed descriptive extension, not as an estimate of an undefined correlation. A question has no defined $\hat{\rho}$ when all its sampled continuations have the same outcome; within-prefix agreement alone is insufficient. The conditional median is not a corpus average. Panel (b) reports, for each budget, the run-level estimate $\overline{Y}$ of succ@k (Proposition 3.4), the per-run classical estimate $\widehat{C}_{\text{run}}$, the pooled classical estimate $\widehat{C}_{\text{pooled}}$ that a pooled pipeline reports, the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ with a 95% interval resampled over questions, and the second-order reference of (6) on the same denominator, that is, averaged over all 200 questions with the variance estimate set to its well-defined value of zero wherever $\hat{\rho}$ is undefined. The $k = 1$ row is the negative control: its interval is reported and contains zero, while its second-order reference cell carries a dash rather than a number because the second-order term vanishes identically at $k = 1$, carrying a factor $\binom{k}{2}$. Panel (c) compares the original model with two 2026 open-weight models (Qwen3.5-9B and granite-4.2-8b, thinking disabled). They share the nominal task and 16×8 design; granite-4.2-8b retains 198 of the requested 200 questions, those that reached their quota of admissible prefixes within a fixed number of redraw rounds. Each $\hat{\rho}$ mean and median uses only that model's defined questions; undefined correlations are not measured zeros. Containment is empirical, against the same held-out reference in every column, not population coverage. Model, generation settings, tokenizer and template vary together. The pooled excess has the same direction in these settings, but its magnitude does not form a monotone gradient with the conditional mean or median correlation.

(a) Within-run correlation across questions

median $\hat{\rho}$, conditional on the 62 questions where defined	0.221
quartiles (Q_1, Q_3), same 62 questions	(0.016, 0.649)
median $\hat{\rho}$ (quartiles), all 200, undefined set to 0	0.000 (0.000, 0.000)
questions on which $\hat{\rho}$ is defined	62 of 200

(b) Pooled excess by budget

k	$\overline{Y}$	$\widehat{C}_{\text{run}}$	$\widehat{C}_{\text{pooled}}$	$\widehat{C}_{\text{pooled}} - \overline{Y}$	95% CI	second-order ref. (6)
1	0.8981	0.8959	0.8959	−0.0023	[−0.0073, +0.0028]	—
2	0.9197	0.9167	0.9320	+0.0123	[+0.0059, +0.0194]	0.0154
4	0.9309	0.9322	0.9560	**+0.0251**	[+0.0155, +0.0367]	0.0192
8	0.9456	0.9456	0.9714	+0.0257	[+0.0143, +0.0393]	0.0164

(c) Common readouts across three models

Quantity	Qwen2.5-7B	Qwen3.5-9B	granite-4.2-8b
Retained questions	200	200	198
Defined $\hat{\rho}$	62	57	63
Undefined $\hat{\rho}$	138	143	135
Conditional mean $\hat{\rho}$	0.3293	0.4068	0.2058
Conditional median $\hat{\rho}$	0.2206	0.3340	0.0028
$\hat{\mu}_1$	0.8959	0.9123	0.8748
$\overline{Y}$, $k = 4$	0.9309	0.9306	0.9075
$\widehat{C}_{\text{pooled}}$, $k = 4$	0.9560	0.9456	0.9186
$\widehat{C}_{\text{pooled}} - \overline{Y}$, $k = 4$	+0.0251	+0.0150	+0.0111
95% paired CI	[+0.0155, +0.0367]	[+0.0087, +0.0225]	[+0.0052, +0.0186]
Attempt-level containment	0.864	0.890	0.893
RUNCS containment	0.966	0.959	0.964
i.i.d. pooled excess, $k = 4$	−0.0004	+0.0005	−0.0005

agrees, their measured gap is exactly zero, and they stay in every denominator below. On 20 of the 62 the gap is exactly zero as well: fewer than $k = 4$ of their prefixes contain a failed continuation and every prefix still succeeds within four, so $\widehat{U}_4 = \overline{Y}_4 = 1$; this is why §6.1 counts 158 zero-gap questions. Table 31 reports a sensitivity calculation under additional mean and homogeneity assumptions; observed agreement alone does not bound their latent gap. The estimand gap follows the

Table 9: On the headline design the estimand gap is positive from $k = 2$ on, within a factor of two of the second-order reference, and the finite-design expectation predicts the pooled excess closely. The gap is estimated by a statistic that is unbiased for the quantity the second-order reference is about (16 prefixes × 8 continuations on 200 questions). $\widehat{U}_k$ is the cross-run U-statistic of Proposition A.8, unbiased for pass@k^{iid} whenever $R \geq k$, which this design satisfies at every budget; $\overline{Y}$ is unbiased for succ@k. Their difference is therefore an unbiased estimate of pass@k^{iid} – succ@k and the second-order formula of (6) applies to it verbatim. The last two columns keep the pooled estimator the literature reports: the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$, and its prediction, the finite-design expectation $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ of Proposition A.21 evaluated at each question's estimated moments minus the same moments' unbiased estimate of succ@k, which is $\widehat{C}_{\text{run}}$; the two terms of the prediction have the expectations of the two terms of $\widehat{C}_{\text{pooled}} - \overline{Y}$. At $R = 16$, $n = 8$ the collision-free fraction of pooled k-subsets is 0.699 at $k = 4$ and 0.151 at $k = 8$, and that is exactly the discrepancy the finite-design prediction carries and the second-order reference does not. Intervals are resampled over questions. The $k = 1$ row is the negative control: no dependence term exists there, both cells carry a dash, and the measured gap's interval contains zero.

k	$\overline{Y}$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}$	95% CI	second-order ref. (6)	$\widehat{C}_{\text{pooled}} - \overline{Y}$	predicted $\mathbb{E}[\widehat{C}_{\text{pooled}}] - \widehat{C}_{\text{run}}$, by (10)
1	0.8981	0.8959	−0.0023	[−0.0071, +0.0027]	—	−0.0023	—
2	0.9197	0.9328	+0.0132	[+0.0067, +0.0205]	0.0154	+0.0123	0.0141
4	0.9309	0.9571	**+0.0262**	[+0.0156, +0.0375]	0.0192	+0.0251	0.0223
8	0.9456	0.9723	+0.0266	[+0.0148, +0.0409]	0.0164	+0.0257	0.0243

Table 10: On the headline design, moving the unit from the attempt to the run does the work (budget $k = 4$, $\delta = 0.05$): at a fixed sample of runs, containment goes from 0.864 to 0.960 and the mean width from 0.028 to 0.402, and the zero-width intervals disappear. Making the run-level interval valid at any stopping time (run-level runCS) is comparatively cheap: a further 0.006 of containment and 0.075 of width. Each question's runs are split into an estimation half and a held-out half; the held-out half supplies the held-out reference, a run-level estimate of succ@k from runs the interval did not see. Columns: interval; containment of the held-out reference; mean width; zero width, the fraction of intervals whose two endpoints coincide. All six rows face the same held-out reference and the same estimation half, so the contrasts between rows are the quantity of interest; the absolute levels are containment of a noisy reference, which has no general conservative direction. The last three rows are further run-level intervals on the 8 run indicators of the estimation half: the run-level percentile bootstrap; the same bootstrap with the $(R' - 1)/R'$ plug-in variance factor of a resampled proportion removed, R' being the number of runs in the estimation half; and the run-level Wald interval on the unbiased variance. They contain the reference 0.957, 0.957 and 0.947 of the time, but a fraction 0.861 of their realisations is a single point: with 8 runs and a saturated question the resampled proportion is constant, and neither the variance factor nor the unbiased variance moves a degenerate interval. That failure mode is counted, not repaired (no continuity correction, no Agresti–Coull, no transformation).

Interval	Containment	Mean width	Zero width
attempt-level bootstrap	0.864	0.028	76.9%
run-level Clopper–Pearson	0.960	0.402	0%
run-level runCS	0.966	0.477	0%
run-level bootstrap	0.957	0.076	86.1%
run-level bootstrap, factor removed	0.957	0.080	86.1%
run-level Wald, unbiased variance	0.947	0.076	86.1%

theory in sign and size, and Table 9 measures it with the statistic the theory is about: with $R = 16 \geq k$ the cross-run U-statistic of Proposition A.8 is unbiased for pass@k^{iid}, so $\widehat{U}_k - \overline{Y}$ is an unbiased estimate of pass@k^{iid} – succ@k. It is a null at $k = 1$, where no dependence term exists, and grows to +0.0262 (interval +0.0156 to +0.0375) at $k = 4$, against a second-order reference of 0.0192 from the same questions; at $k = 8$ it is +0.0266 against 0.0164, the reference falling behind because the second-order truncation discards the higher moments that a budget of eight begins to weight. The pooled estimator that a pooled pipeline reports tells the same story once it is read against the right quantity: Table 8(b) gives $\widehat{C}_{\text{pooled}} - \overline{Y} = +0.0251$ at $k = 4$, and the finite-design expectation of Proposition A.21, evaluated with unbiased moment estimates from the same runs, predicts 0.0223, a ratio of 1.12; across the three budgets the ratio stays within 0.87–1.12, and the residual has the direction and order of the

plug-in's own bias (Appendix C.2); fitted on one half of each question's runs and scored on the other, the same expectation reproduces the held-out pooled number to within 1.003× at $k = 4$ (Table 30). The containment comparison is the practical content (Table 10). An attempt-level bootstrap that pools the continuations of a question contains the held-out reference 86% of the time with a mean width of 0.03, and three quarters of its intervals have width zero, because pooling makes the count of successes look like a large binomial sample; runCS contains it 97% with a width of 0.48, which is the width eight runs support; with the question as the unit, the two containment figures carry question-cluster bootstrap intervals of $[0.822, 0.906]$ and $[0.950, 0.980]$ (Table 34), two to six times wider than a binomial standard error over the 50 splits would suggest. Both figures are measured against a held-out estimate that carries its own noise, and Table 32 calibrates what that costs: replaying the protocol on laws fitted question by question to these data reproduces the measured 0.86 and 0.97 against the held-out reference, and against the exact succ@k of the same laws the two become 0.768 and 1.000 — in those fitted laws the noisy held-out reference raises attempt-level containment and lowers run-level containment. Table 33 is the same calibration for the larger design of §6.3, and the two corrections differ, so neither transfers to the other's design. The i.i.d. control, built by cutting pseudo-prefixes into the same model's independent chains, shows no correlation and no gap under either statistic, so the method does not report dependence where there is none; it is evaluated on the first 100 of these 200 questions, its median $\hat{\rho}$ over the 32 of those on which the correlation is defined against 62 of 200 here, so it is a common-subset diagnostic and not a control on the main denominator (Table 12).

The second-order reference column of Table 8(b) is the second-order approximation to pass@k^{iid} − succ@k, not to $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ − succ@k: at $R = 16$ and $n = 8$ about 30% of the pooled 4-subsets fall twice inside one prefix, which by Corollary A.23 places $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ below pass@k^{iid} by an amount that column does not carry and that the finite-design column of Table 9 does. The comparison with it is therefore a check on the magnitude of the gap — measured and reference agree to within a factor between 0.85 and 1.63 across the budgets of Table 9 and Table 17 — and not a fitted correction; the finite-design expectation is the fitted quantity, and it is exact up to the moment estimates it is fed.

Two robustness checks of §6.1. A fresh-prefix resample of 30 of the agreeing questions leaves 4 with a definable correlation (Table S3), so the agreement is a property of the questions and not of the draw; and the same design with no admissibility condition moves the $k = 4$ gap by +0.0011, with paired interval $[-0.0059, +0.0081]$ (Table 11).

The supplementary report (Supplement §S2.1) restricts the gap to the questions with a defined correlation, repeats the fresh-prefix resample in full, and rescores the intervals against Arm F. Arm F is a supplementary batch on the headline design's model that redraws 128 fresh prefixes, of $n = 8$ continuations each, on the 62 questions whose correlation is defined, with the same prompt, sampling parameters and verifier; its 128 runs give a fresh reference.

B.3 Replication across pretraining families

The shared-prefix design is repeated, unchanged, on 5 further open-weight models — Qwen3.5-9B, granite-4.2-8b, Llama-3.1-8B, OLMo-3-7B and gemma-4-E4B — each with its own independent-chain control generated in the same batch. Table 13 puts all 6 runs (the headline design and the 5 further models), spanning 5 pretraining families, in one grid, each cell computed from its own run. The three-part direction rule — (a) the independent-chain control has $|\hat{\rho}|$ median at most 0.05, (b) the shared-prefix run has $\hat{\rho}$ median above zero, and (c) the $k = 4$ pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ has positive mean and a question-bootstrap lower bound above zero — holds in every one of the 6 rows. Two quantities are reported for each model at $k = 4$ (Table 14): the estimand gap $\widehat{U}_4 - \overline{Y}$, which is +0.0154, +0.0117, +0.0609, +0.0211 and +0.0110 on the five models in the order listed above, and the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$, which is +0.0150, +0.0111, +0.0585, +0.0203 and +0.0103. Each has a question-bootstrap interval excluding zero on every model, against independent-chain controls whose median correlation rounds to zero.

Two 2026 open-weight models. The headline design (16 × 8) is repeated on Qwen3.5-9B and granite-4.2-8b, with thinking disabled and independent-chain controls generated separately for each model. The former retains 200 questions; the latter retains 198, the questions that reached their quota of admissible prefixes within a fixed number of redraw rounds. For granite-4.2-8b, $\widehat{C}_{\text{pooled}} = 0.9186$

Table 11: Removing the prefix admissibility condition leaves the headline gap unchanged within its interval (paired $k = 4$ difference +0.0011, interval $[-0.0059, +0.0081]$). The same 200 questions, the same model, prompt, sampling parameters, verifier, prefix length and design cell as the headline design, with no admissibility condition on the prefix: prefixes that already contain an answer and prefixes that terminate within the first P tokens are kept, and a run that terminates inside its prefix is scored on the answer it has already produced. The two protocols estimate two different quantities, so the difference is a change of estimand, not a bias. Intervals are percentile bootstrap over questions.

(a) paired difference, unconditional minus conditional

	difference [95% CI]	questions
$\hat{\rho}$	+0.0185 [−0.0490, +0.0924]	53
$\hat{\mu}_1$	−0.0011 [−0.0063, +0.0039]	200
$\hat{U}_4 - \bar{Y}$	+0.0011 [−0.0059, +0.0081]	200

(b) the gap under each prefix protocol, all questions

	unconditional	conditional (headline)
$\hat{U}_1 - \bar{Y}$	+0.0007 [−0.0048, +0.0065]	−0.0023 [−0.0068, +0.0029]
$\hat{U}_2 - \bar{Y}$	+0.0148 [+0.0075, +0.0227]	+0.0132 [+0.0067, +0.0203]
$\hat{U}_4 - \bar{Y}$	+0.0273 [+0.0163, +0.0395]	+0.0262 [+0.0160, +0.0375]
$\hat{U}_8 - \bar{Y}$	+0.0307 [+0.0180, +0.0444]	+0.0266 [+0.0144, +0.0410]

(c) what the admissibility condition removes

	prefixes	share
prefix already carries an answer	20	0.006
prefix terminates within P tokens	20	0.006
neither (the headline design's admissible class)	3,160	0.988
removed by the headline design's condition	40	0.0125
same share, per question, in the headline design's own log	n/a	0.0122

Table 12: With no shared state the pipeline reports neither dependence nor a gap: the median correlation is near zero and every pooled-excess interval contains zero. The same analysis pipeline is applied to sampling regimes with no shared random intermediate, where $\mathrm{Var}(\theta) = 0$ and therefore $\mathrm{succ@}k = \mathrm{pass@}k^{\mathrm{iid}}$. Panel (a) uses independently decoded chains, cut into pseudo-runs of the same shape ($R = 16$ runs of $n = 8$ attempts, 20 groupings), for the first 100 of the 200 questions of the shared-prefix experiment; its question set is a proper subset of that experiment's and its denominators are its own, so this panel is a diagnostic on a common subset and not a control evaluated on the full main-analysis denominator. Its pooled-excess row is an average over all 100 of those questions, while its median $\hat{\rho}$ is over the 32 of them on which the within-run correlation is defined — against 62 of 200 in the main analysis — so the two medians are not computed on the same question set either. Panel (b) reshuffles the sibling groups of the reasoning tree so that the shared prefix is destroyed, and its pooled-excess entry is the range over the three estimable budgets, each of whose intervals contains zero. In both, the within-run correlation should sit at zero and the pooled-excess interval should contain zero; it does. The correlation is a property of a run, not of a budget, and the runs summarise it by quartiles rather than by an interval, so the budget and containment cells of those two rows are n/a.

Regime	Quantity	k	Measured	Contains zero
(a) independent chains	median $\hat{\rho}$	n/a	−0.0016	n/a
(a) independent chains	$\widehat{C}_{\mathrm{pooled}} - \overline{Y}$	4	−0.0004	yes
(b) reshuffled tree groups	median $\hat{\rho}$	n/a	−0.0070	n/a
(b) reshuffled tree groups	$\widehat{C}_{\mathrm{pooled}} - \overline{Y}$	1, 2, 4	−0.0009 to +0.0004	yes

and $\overline{Y} = 0.9075$ at $k = 4$, a pooled excess $\widehat{C}_{\mathrm{pooled}} - \overline{Y}$ of +0.0111 with question-bootstrap interval $[+0.0052, +0.0186]$; its estimand gap $\widehat{U}_4 - \overline{Y}$ is +0.0117 (Table 14). Its correlation has conditional mean 0.2058 and median 0.0028 on the 63 defined questions. Table 8(c) uses the held-out reference for all three containment columns: Granite's attempt-level and run-level empirical containments are

0.893 and 0.964. Its independent-chain control has a $k = 4$ pooled excess of −0.0005. On Qwen3.5-9B the $k = 4$ pooled excess is +0.0150 [+0.0087, +0.0225] and the estimand gap +0.0154 (Table 14).

Denominators, cell by cell. Table 13 does not put every row on one denominator, and Table 15 prints the denominator of each. The pooled-excess row averages over every retained question. The second-order reference of (6) is a function of the within-run correlation, which is undefined on a question whose runs all agree, so the field the table selects is the average over the defined-$\hat{\rho}$ questions alone — 62 of 200 in the Qwen2.5 column. That is why its 0.0620 sits beside a whole-corpus pooled excess of +0.0251 and looks inconsistent with the 0.0192 of Table 9, which is the same quantity zero-filled over all 200 questions: an agreeing question has a latent variance estimate of zero and therefore a second-order term of exactly zero, and zero-filled × retained = conditional × defined holds to machine precision. Recomputed that way the six columns read 0.0086 to 0.0431, and the Qwen2.5 entry is 0.0192, reproducing Table 9's cell. Neither reading is comparable across columns, because the defined subset differs from column to column.

What this block does not license. The definable-$\hat{\rho}$ denominators differ widely across columns — from 46 to 127 of about 200 — so the conditional correlations are not comparable between columns, and an undefined correlation is not a measured zero. Model, generation settings, tokenizer and chat template vary together, so no causal comparison between families is available and none is claimed; what replicates is the sign of the pooled excess and the direction rule. One further family was attempted and not run, because the generation code could not load its released checkpoint.

Table 13: The headline design repeated on five further open-weight models: the three-part direction rule of Appendix B.3 holds on all 6 runs, which span 5 pretraining families (the two Qwen columns are two generations of one family). The column headers abbreviate `Qwen2.5-7B-Instruct`, `Qwen3.5-9B`, `granite-4.2-8b`, `Llama-3.1-8B-Instruct`, `Olmo-3-7B-Instruct` and `gemma-4-E4B-it`; the *Model* row repeats them in the form the body uses. Each cell is computed from its own run alone; nothing is pooled or weighted across runs. The $\widehat{C}_{\text{pooled}} - \overline{Y}$ rows are the pooled excess; the estimand gap $\widehat{U}_4 - \overline{Y}$ of the same runs is in Table 14. Each column keeps its own denominators: $\hat{\rho}$ is undefined on a question whose prefixes all agree; such questions have a pooled excess of exactly zero and stay in its denominator, while the $\hat{\rho}$ rows use the "defined" count of their own column, which ranges from 46 to 127 of about 200. *Undefined correlations are not measured zeros, and the conditional correlations are not comparable across columns.* The second-order reference row is averaged over those defined-$\hat{\rho}$ questions too, so it is not the zero-filled reading that Table 9 and §6.1 quote for the same run (0.0620 here against 0.0192 there in the Qwen2.5 column); Table 15 prints both readings and the exact identity between them. Intervals are 95% bootstrap intervals resampled over questions, per cell, with no multiplicity adjustment. Containment is of the held-out reference. Model, generation settings, tokenizer and chat template vary together across columns, so no causal reading of the differences between columns is offered: the reading is the *sign* of the pooled excess and the direction rule.

Reading	Qwen2.5	Qwen3.5	granite	Llama	OLMo	gemma
Pretraining family	Qwen	Qwen	Granite	Llama	OLMo	Gemma
Model	Qwen2.5-7B	Qwen3.5-9B	granite-4.2-8b	Llama-3.1-8B	OLMo-3-7B	gemma-4-E4B
Retained questions	200	200	198	198	187	200
Defined $\hat{\rho}$	62	57	63	127	62	46
Undefined $\hat{\rho}$	138	143	135	71	125	154
Conditional median $\hat{\rho}$	0.2206	0.3340	0.0028	0.4286	0.2698	0.1732
Conditional mean $\hat{\rho}$	0.3293	0.4068	0.2058	0.4538	0.3747	0.2741
95% CI, lower	0.2487	0.3154	0.1285	0.3920	0.2833	0.1902
95% CI, upper	0.4147	0.4937	0.2875	0.5113	0.4681	0.3703
$\hat{\mu}_1$	0.8959	0.9123	0.8748	0.8087	0.8920	0.9100
95% CI, lower	0.8603	0.8738	0.8335	0.7706	0.8545	0.8736
95% CI, upper	0.9276	0.9441	0.9118	0.8440	0.9251	0.9432
$\widehat{C}_{\text{pooled}} - \overline{Y}$, $k = 4$	**+0.0251**	**+0.0150**	**+0.0111**	**+0.0585**	**+0.0203**	**+0.0103**
95% CI, lower	+0.0155	+0.0087	+0.0052	+0.0445	+0.0115	+0.0038
95% CI, upper	+0.0367	+0.0225	+0.0186	+0.0735	+0.0303	+0.0182

continued on the next page

Table 13, continued.

Reading	Qwen2.5	Qwen3.5	granite	Llama	OLMo	gemma
Second-order ref. (6), defined-$\hat{\rho}$ questions	0.0620	0.0303	0.0311	0.0672	0.0418	0.0464
i.i.d. control median $\hat{\rho}$	−0.0016	0.0000	0.0000	0.0000	0.0000	0.0000
Attempt-level containment, held-out	0.864	0.890	0.893	0.649	0.862	0.902
RUNCS containment, held-out	0.966	0.959	0.964	0.899	0.956	0.966
Direction rule (a)–(c)	YES	YES	YES	YES	YES	YES

Table 14: On each of the five further models the estimand gap and the pooled excess at $k = 4$ are positive, with question-bootstrap intervals above zero. The shared-prefix design of §6.1 repeated unchanged on five further open-weight models, with the headline design in the last row ($R = 16$ prefixes, $n = 8$ continuations, $k = 4$). $\widehat{U}_4 - \overline{Y}$ is the estimand gap of Proposition A.8, $\widehat{C}_{\text{pooled}} - \overline{Y}$ the pooled excess, each a question average with a 95% percentile question-bootstrap interval (2,000 resamples, no multiplicity adjustment). Release is the model's public release (month only where only the month is recorded). The last column counts questions on which every continuation of every prefix agrees; their gap is identically zero and they stay in every denominator. *Run with the chat template's thinking switch off (`enable_thinking`, `thinking`). Model, tokenizer and chat template vary together across rows, so differences between rows have no causal reading.

Model	release	Q	$\widehat{U}_4 - \overline{Y}$ [95% CI]	$\widehat{C}_{\text{pooled}} - \overline{Y}$ [95% CI]	all agree
`Qwen3.5-9B`*	2026-03-02	200	+0.0154 [+0.0088, +0.0231]	+0.0150 [+0.0088, +0.0222]	143
`granite-4.2-8b`*	2026-08-25	198	+0.0117 [+0.0052, +0.0198]	+0.0111 [+0.0052, +0.0180]	135
`gemma-4-E4B`*	2026-04-02	200	+0.0110 [+0.0044, +0.0195]	+0.0103 [+0.0040, +0.0181]	154
`OLMo-3-7B`	2025-11-20	187	+0.0211 [+0.0123, +0.0315]	+0.0203 [+0.0113, +0.0302]	125
`Llama-3.1-8B`	2024-07	198	+0.0609 [+0.0463, +0.0768]	+0.0585 [+0.0444, +0.0732]	71
`Qwen2.5-7B` (headline design)	2024-09	200	+0.0262 [+0.0161, +0.0373]	+0.0251 [+0.0154, +0.0368]	138

Table 15: The second-order reference of Table 13 averages over the questions with a defined correlation; zero-filled over all retained questions it reproduces the value of Table 9. The denominator of every $k = 4$ cell of Table 13, and the whole-corpus reading of its second-order reference, recomputed from the six runs' per-question records. The pooled-excess row averages over every retained question. The second-order reference of (6) is a function of the within-run correlation, which is undefined on a question whose runs all agree, so the recorded field averages over the defined-$\hat{\rho}$ questions only: the two rows do not share a denominator, and Table 13 prints the conditional one. The zero-filled row is the same quantity averaged over all retained questions with the well-defined value 0 substituted on the agreeing questions — their latent variance estimate is zero, so their second-order term is zero — which is the convention Table 9 uses. The two readings are one measurement under two denominators: zero-filled × retained = conditional × defined, exactly, and the Qwen2.5 zero-filled value reproduces Table 9's $k = 4$ reference cell to machine precision. Neither reading is comparable across columns, because the defined subset differs from column to column. Intervals are percentile bootstrap over questions, 2000 resamples.

Reading	Qwen2.5	Qwen3.5	granite	Llama	OLMo	gemma
Retained questions	200	200	198	198	187	200
Defined $\hat{\rho}$	62	57	63	127	62	46
Undefined $\hat{\rho}$	138	143	135	71	125	154
$\widehat{C}_{\text{pooled}} - \overline{Y}$, $k = 4$	+0.0251	+0.0150	+0.0111	+0.0585	+0.0203	+0.0103
denominator	200	200	198	198	187	200
Second-order ref., conditional	0.0620	0.0303	0.0311	0.0672	0.0418	0.0464
denominator (defined $\hat{\rho}$)	62	57	63	127	62	46
Second-order ref., zero-filled	0.0192	0.0086	0.0099	0.0431	0.0139	0.0107
95% CI, lower	0.0104	0.0033	0.0035	0.0310	0.0066	0.0033
95% CI, upper	0.0294	0.0150	0.0188	0.0580	0.0222	0.0198
denominator (retained)	200	200	198	198	187	200

B.4 A PUBLIC REASONING TREE

This section is the complete version of Section 6.2 of the main text.

Table 16: On the public reasoning tree the pooled excess grows with the budget as the second-order reference predicts, and run-level RUNCS contains the held-out reference more often than the attempt-level bootstrap. The corpus is a process-supervision corpus that we did not generate (PRM800K), in which a group of sibling continuations of one annotated prefix is a run, a sibling is an attempt and the verifier is a human step-level label. Columns are as in Table 8(b), with the second-order reference of (6) on the same denominator; budgets stop at the group size ($k \le 4$). Siblings are only approximately exchangeable — the first fresh sibling at a node succeeds about two percentage points more often than the later ones — so the classical estimator targets the subset average of Proposition 3.3, and $\widehat{C}_{\text{pooled}} - \widehat{C}_{\text{run}}$ is the contrast against that average. At $k = 4$ a group holds exactly four siblings, so $\widehat{C}_{\text{run}}$ coincides with $\overline{Y}$; at $k = 2$, $\widehat{C}_{\text{run}} - \overline{Y} = -0.0068$, the whole footprint of the position effect, small beside the $k = 2$ pooled excess of +0.0506. Intervals are resampled over questions; n/a marks a quantity that does not exist here: containment is a single number, and the $\widehat{C}_{\text{pooled}} - \widehat{C}_{\text{run}}$ rows share the reference of the $\widehat{C}_{\text{pooled}} - \overline{Y}$ row at the same budget.

(a) Reasoning tree: attempt-level structure

questions / sibling groups				8,061 / 47,845
median $\hat{\rho}$ (quartiles)				0.289 (0.039, 0.467)
Quantity	k	measured	95% CI	second-order ref. (6)
$\widehat{C}_{\text{pooled}} - \overline{Y}$	2	+0.0506	[+0.0473, +0.0540]	0.0577
$\widehat{C}_{\text{pooled}} - \overline{Y}$	4	**+0.1276**	[+0.1245, +0.1309]	0.1267
$\widehat{C}_{\text{pooled}} - \overline{Y}$	8	not estimable at the available group size		
$\widehat{C}_{\text{pooled}} - \widehat{C}_{\text{run}}$	2	+0.0574	[+0.0559, +0.0589]	n/a
$\widehat{C}_{\text{pooled}} - \widehat{C}_{\text{run}}$	4	+0.1276	[+0.1242, +0.1308]	n/a
Containment, attempt-level bootstrap	4	0.609	n/a	n/a
Containment, run-level RUNCS	4	0.788	n/a	n/a

What the released corpus fixes about the unit, and what it does not. Three facts about the sampling unit are checkable in the released files and one is not, and we separate them. *Checkable.* (i) In the stratum that carries these numbers each annotated trajectory contributes exactly one sibling group: 47,845 groups over 47,845 distinct trajectory identifiers, a maximum of one group per trajectory, so no two groups here are nested inside a shared trajectory. (ii) The 8,061 questions carry a median of 4 trajectories each (quartiles 3 and 8), and the question is the level the intervals resample over. (iii) The siblings of a group are the completions the annotation process drew at one node, and the corpus records their order, which is what makes the position analysis below possible. *Not checkable from the release.* Whether two trajectories on the same question were generated by the same model checkpoint, in the same batch, or under a selection rule that depends on earlier trajectories — the released files carry no batch identifier, no sampling timestamp and no record of the selection policy, so nothing in them establishes that trajectories are independent and identically distributed given the question. We therefore state it as a working assumption rather than a property of the data: *trajectories within a question are treated as independent repetitions of one generating procedure*, in the sense of Assumption 5.1, and every interval in this subsection is valid only under it. The assumption is falsifiable in the direction that matters — shared state between trajectories would inflate the between-trajectory agreement — and the one-group-per-trajectory design of (i) removes the nesting we *can* see. It does not remove what we cannot. The secondary stratum used for the sensitivity analysis of Table S6 *is* nested, with a median of 5 groups per trajectory, and is kept separate for that reason.

Which sibling defines the i.i.d. counterfactual. $\widehat{U}_k$ is built from the success count of every sibling in a group, so it is invariant to sibling order, and, because distinct groups are independent, it is unbiased for $1 - (1 - p^{(\text{all})})^k$ with $p^{(\text{all})}$ the all-position marginal — the budget-k probability of a procedure that draws a uniformly random sibling in each of k independent groups. That factorisation uses independence across groups only, not exchangeability within one, which is exactly the property the position effect violates. Reading the first sibling of each group instead gives an equally unbiased estimator $\widehat{U}_k^{(1)}$ of the first-sibling counterfactual $1 - (1 - p^{(1)})^k$. Table 18 reports both on the same

Table 17: On the reasoning tree the estimand gap is positive at both budgets in all three strata and within a quarter of the second-order reference. The reanalysis of Table 9, on the reasoning tree. Here the number of runs per question is not fixed by design but by how many times the annotation process happened to branch, so $\widehat{U}_k$ exists only on the questions carrying at least k sibling groups; the third column names that subset and columns 4–7 are all computed on it, so they share a denominator. The last column is the whole-corpus pooled excess that Table 16 reports, printed here so the two denominators are never read against each other by accident: at $k = 4$ the subset is roughly three fifths of the corpus and it is the harder, more branched part of it. The first block is the main corpus, the second a held-out split of unseen questions, the third an earlier annotation phase with six siblings per group.

Stratum	k	questions, $R \geq k$	$\widehat{U}_k - \overline{Y}$	95% CI	second-order ref. (6)	$\widehat{C}_{\text{pooled}} - \overline{Y}$ (same subset)	$\widehat{C}_{\text{pooled}} - \overline{Y}$ (whole corpus)
main corpus	2	8,061	+0.0671	[+0.0634, +0.0709]	0.0577	+0.0506	+0.0506
	4	4,880	+0.1668	[+0.1628, +0.1707]	0.1530	+0.1464	+0.1276
held-out split	2	403	+0.0726	[+0.0554, +0.0892]	0.0586	+0.0554	+0.0554
	4	233	+0.1642	[+0.1442, +0.1847]	0.1357	+0.1362	+0.1236
earlier phase	2	583	+0.0558	[+0.0411, +0.0701]	0.0549	+0.0418	+0.0418
	4	390	+0.1694	[+0.1530, +0.1869]	0.1648	+0.1440	+0.1188

question set $Q_k = \{q : R_q \geq k\}$ the main estimate uses, with the same per-question weights and against the same $\overline{Y}$: at $k = 4$ the gap is +0.1799 [+0.1728, +0.1871] under the first-sibling reading and +0.1668 under the uniform one, which is the main estimate, a paired difference of +0.0130 [+0.0076, +0.0186]; at $k = 2$ the three figures are +0.0868, +0.0671 and +0.0197. The conclusion does not turn on which position names the target.

What the tree shows. The reasoning tree is a second reading of (6) on a different mechanism, and a weaker one than the arms we generate: its siblings are not exchangeable within a group, as the position effect below shows, so the numbers are consistent with the latent-mixture identity rather than a direct verification of it. Success here is an annotator's label on a sibling completion at a selected step, not end-to-end task success. At budgets $k \leq 4$, which is the group size, the readings are these: the within-run correlation has median 0.29, the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ at $k = 4$ is +0.128 against a second-order prediction of 0.127, and on the questions carrying at least four sibling groups, where the cross-run statistic exists, $\widehat{U}_4 - \overline{Y}$ is +0.1668 (interval +0.1628 to +0.1707) against a reference of 0.153 (Table 17; that subset is the more branched, harder part of the corpus, and the table prints both denominators). The gap is six to seven times the shared-prefix value on the $\widehat{U}_k$ reading (five to six on the pooled excess), in the direction the formula makes explicit — the step-level success rate is 0.38 rather than 0.90, and the factor $(1 - \mu_1)^{k-2}$ is correspondingly larger.

Containment on the tree. Attempt-level intervals contain the reference 61% of the time where runCS contains it 79%; both figures are scored against a held-out reference that here holds as few as two groups, and Table 35 replays the protocol on laws fitted to this corpus and scores it against the exact succ@k: at $k = 4$ runCS then covers 0.998 and the attempt-level interval 0.640, so reference noise explains the run-level shortfall within those fitted laws.

Position effect and end-to-end stratum. The same corpus also shows where the model's exchangeability assumption is only approximate: the first fresh sibling at each node succeeds at a rate 0.0134 higher than the all-position rate, pooled over all 49,266 groups, a position effect large enough to be significant at this scale, so the classical estimator there targets the subset average of Proposition 3.3 rather than the leading-k probability. The position effect does not bound the reported $\hat{\rho}$ in either direction; what it changes is which marginal the i.i.d. counterfactual is built on, and Table 18 measures that directly. A small stratum of the same corpus in which every sibling carries a final answer scored against the reference gives an end-to-end reading of the same mechanism, with a pooled excess of +0.269 at $k = 4$ and a corpus-level correlation of 0.585 (Table S6 in the supplementary report).

Table 18: The reasoning-tree gap is positive under either reading of the i.i.d. counterfactual, first sibling or uniform sibling, and the two readings differ by less than a third of the gap. The position effect is measured on the same question set, the same per-question weights and the same estimator as the main estimate $\widehat{U}_k - \overline{Y}$ of Table 17; the all-position column equals that table's figures. Panel (a): both readings of the i.i.d. counterfactual, each estimated without bias on the subset $Q_k = \{q : R_q \geq k\}$ that the main estimate uses, with questions weighted equally in both. $\widehat{U}_k^{(\mathrm{all})}$ is built from each group's success count over all 4 siblings and is unbiased for $1 - (1 - p_q^{(\mathrm{all})})^k$, the budget-$k$ probability of a procedure that draws a uniformly random sibling in each of k independent groups; $\widehat{U}_k^{(1)}$ is built from the first sibling of each group alone and is unbiased for $1 - (1 - p_q^{(1)})^k$. Both are unbiased because distinct groups are independent, so the product over the k groups factorises; neither needs within-group exchangeability, which is exactly what the position effect violates. $\overline{Y}$ is the same in both columns. Panel (b): the per-question marginals behind them, the matched mean-value bound $B_k^{\mathrm{match}} = (k/|Q_k|) \sum_q |\hat{p}_q^{(1)} - \hat{p}_q^{(\mathrm{all})}|$ and the per-question plug-in difference $D_k = |Q_k|^{-1} \sum_q [g_k(\hat{p}_q^{(1)}) - g_k(\hat{p}_q^{(\mathrm{all})})]$ it bounds. B_k^{match} is not informative at this scale — the per-question marginal differences do not cancel the way the corpus-pooled ones do — which is why panel (a) estimates the two targets directly instead of bounding their difference. D_k is negative where the unbiased difference is positive because g_k is concave and $\hat{p}_q^{(1)}$, read off R_q siblings rather than $4R_q$, carries the larger sampling variance; that is a property of the plug-in, not of the two targets. Panel (c) gives the corpus-pooled figures for comparison: they bound the difference of two *global* plug-ins, on all 49,266 groups and with no per-question weighting, and therefore do not bound the per-question average of panel (a). Intervals are percentile bootstrap over questions, 2000 resamples, with all columns of a row resampled together so the differences are paired.

Reading	$k = 2$	95% CI	$k = 4$	95% CI
(a) Both readings of the i.i.d. counterfactual, on $Q_k = \{q : R_q \geq k\}$ (8,061 questions at $k = 2$, 4,880 at $k = 4$)				
$\overline{Y}$	0.5421	[0.5365, 0.5480]	0.6323	[0.6267, 0.6383]
$\widehat{U}_k^{(\mathrm{all})}$, uniform sibling	0.6092	[0.6040, 0.6146]	0.7992	[0.7939, 0.8044]
$\widehat{U}_k^{(1)}$, first sibling	0.6289	[0.6214, 0.6363]	0.8122	[0.8047, 0.8198]
$\widehat{U}_k^{(\mathrm{all})} - \overline{Y}$	+0.0671	[+0.0633, +0.0707]	+0.1668	[+0.1628, +0.1710]
$\widehat{U}_k^{(1)} - \overline{Y}$	+0.0868	[+0.0817, +0.0918]	+0.1799	[+0.1728, +0.1871]
$\widehat{U}_k^{(1)} - \widehat{U}_k^{(\mathrm{all})}$	+0.0197	[+0.0149, +0.0245]	+0.0130	[+0.0076, +0.0186]
(b) Per-question marginals and the matched bound, same denominators				

continued on the next page

Table 18, continued.

Reading	$k = 2$	95% CI	$k = 4$	95% CI
$\hat{p}_q^{(\text{all})}$, mean	0.3834	[0.3791, 0.3880]	0.3446	[0.3401, 0.3491]
$\hat{p}_q^{(1)}$, mean	0.4015	[0.3956, 0.4075]	0.3588	[0.3528, 0.3646]
$\hat{p}_q^{(1)} - \hat{p}_q^{(\text{all})}$, mean	+0.0180	[+0.0143, +0.0217]	+0.0142	[+0.0106, +0.0178]
$\lvert\hat{p}_q^{(1)} - \hat{p}_q^{(\text{all})}\rvert$, mean	0.1270	[0.1245, 0.1294]	0.0986	[0.0964, 0.1008]
D_k, per-question plug-in diff.	−0.0107	—	−0.0262	—
B_k^{match}, matched bound	0.2539	—	0.3946	—
(c) The corpus-pooled figures, for contrast: a bound on the difference of two *global* plug-ins, over all 49,266 groups				
$\hat{\mu}^{(1)}$, pooled	0.3641	—	0.3641	—
$\hat{\mu}^{(\text{all})}$, pooled	0.3507	—	0.3507	—
$\Delta\hat{\mu} = \hat{\mu}^{(1)} - \hat{\mu}^{(\text{all})}$, pooled	+0.0134	—	+0.0134	—
$B_k^{\text{pool}} = k\,\lvert\Delta\hat{\mu}\rvert$	0.0268	—	0.0536	—
global plug-in difference	+0.0172	—	+0.0142	—

B.5 Away from saturation

This section is the complete version of Section 6.3 of the main text.

Relative to the headline design, three things change on Arm A and one does not. The dependence is stronger: the smaller model locks onto its prefix, and the within-run correlation has median 0.663 where the larger model's was 0.22. The gap is correspondingly larger and keeps growing with the budget: $\widehat{U}_k - \overline{Y}$ is +0.0976 at $k = 2$, +0.1678 at $k = 4$ and +0.2410 at $k = 32$, with every interval excluding zero and the $k = 1$ control containing it. The second-order reference is now visibly what it is, a leading term: it matches the measured gap at $k = 2$ (1.03×) and falls behind by 1.19–1.24× at $k \geq 4$, in the direction the discarded higher moments account for, and Table S7 shows the reason band by band. What does not change is the finite-design expectation: read against Proposition A.21, the pooled number a pooled pipeline reports is reproduced to within 1.01–1.04× at all five budgets $k \geq 2$, through a regime in which the collision-free fraction of pooled subsets falls from 0.828 at $k = 4$ to 0.012 at $k = 16$ and to below 0.001 at $k = 32$, where essentially every pooled 32-subset draws several continuations from one prefix. The residual again has the direction and order of the plug-in's own bias, and the out-of-sample check of Table 30 reproduces the held-out pooled number at $k = 4$ to within 1.011× on this arm and 1.002× on the next.

The containment comparison is starker than at saturation (Table 20). At $k = 4$ the attempt-level bootstrap contains the held-out reference 0.146 of the time with a mean width of 0.052, where runCS contains it 0.892 with a width of 0.502 and the run-level Clopper–Pearson interval 0.838; at $k = 16$ the attempt-level figure is 0.188 and between a fifth and about a half of its intervals have collapsed to a point, the exact fraction depending on the tolerance at which two endpoints near one count as equal (Table 20). The held-out reference is a sixteen-run estimate and carries noise of its own, and Table 33 calibrates it on *this* design — the same protocol run on laws fitted question by question and scored against the exact succ@k those laws define — rather than on the smaller design of Table 32, whose correction does not transfer. The correction (coverage of the exact succ@k minus containment of the held-out reference, on the fitted laws) does not rescue the attempt-level interval but deepens its failure: it is −0.060 at $k = 4$ and −0.114 at $k = 16$ for the attempt-level bootstrap against +0.099 and +0.102 for runCS, so the noisy reference has been flattering the pooled interval and understating the run-level one. Two quantities follow, and they differ. On the fitted laws themselves the attempt-level bootstrap covers the exact succ@k in 0.135 of trials at $k = 4$ (Table 33; the fitted-law column of Table 22); adding the correction to the containment measured on the real questions instead gives about 0.09 at $k = 4$ and 0.07 at $k = 16$, against about 0.99 for runCS, if the fitted laws are right. Supplement §S2.3 stratifies both arms by difficulty band and by how many prefix draws the admissibility condition discarded.

Arm B, entered from the other side. The mechanism turns out not to need strong dependence. The within-run correlation here is an order of magnitude below Arm A's — median 0.075 (quartiles 0.021 and 0.211) against 0.663 — at $\hat{\mu}_1 = 0.7013$, and the gap is correspondingly smaller but positive at every budget: $\widehat{U}_k - \overline{Y}$ is +0.0360 (interval +0.0174 to +0.0581) at $k = 4$ and +0.0287 at $k = 32$, with the $k = 1$ control's interval only just containing zero (+0.0158, interval −0.0002 to +0.0336). The finite-design expectation tracks it as tightly as on Arm A: read against Proposition A.21 the pooled number is reproduced to within 0.95–1.01× at all five budgets $k \geq 2$. The second-order reference does not, and Corollary A.5 says why: it matches at $k = 2$ (0.97×) and then falls behind, by 1.20× at $k = 4$ and 1.93–4.48× from $k = 8$ on, because at $\hat{\mu}_1 = 0.7013$ the factor $(1 - \mu_1)^{k-2}$ shrinks the leading term faster than it shrinks the higher central moments the truncation discards. That is a failure of the reference, not a discrepancy in the data: the exact expansion has no free parameter and the finite-design expectation, which uses it in full, is tight at the same budgets.

The containment result does not soften with the weaker dependence. At $k = 4$ the attempt-level bootstrap contains the held-out reference 0.208 of the time against 0.922 for runCS and 0.902 for the run-level Clopper–Pearson interval; at $k = 16$ it is 0.463 against 0.962 and 0.947. The Arm B block of Table 33 reports two quantities. On the fitted laws, the attempt-level bootstrap covers the exact succ@k in 0.269 and 0.187 of trials at $k = 4$ and $k = 16$, while runCS reaches 1.000 at both budgets; the correction, that coverage minus containment of the held-out reference, is +0.047 and −0.279 for the attempt-level bootstrap. The noise thus moves the attempt-level number in either direction depending on the budget, and never far enough to matter. Read together the two arms make a point neither makes alone: an order of magnitude of within-run correlation separates them, and the

gap is positive at every budget in both, the finite-design expectation is tight in both, and attempt-level containment collapses in both.

Table 19: Away from saturation the estimand gap is positive at every budget $k \geq 2$ on each arm, with an interval above zero; the second-order reference falls behind it from $k = 4$ on, while the finite-design expectation keeps predicting the pooled excess. Within-run dependence and the estimand gap in a regime where neither the second-order term nor the U-statistic is squeezed by the design. Arm A is GSM8K answered by a 1.5B instruction-tuned model: 194 questions carrying $R = 32$ independently drawn reasoning prefixes $\times\, n = 32$ continuations each, exact answer match as verifier. Because $R \geq k$ at every budget on the grid, the cross-run U-statistic $\widehat{U}_k$ of Proposition A.8, unbiased for pass@k^{iid}, is available up to $k = 32$; $\overline{Y}$ is unbiased for succ@k (Proposition 3.4), so their difference is an unbiased estimate of pass@k^{iid} − succ@k and the second-order formula of (6) applies to it verbatim, on the same denominator, that is, averaged over all questions with the variance estimate set to its well-defined value of zero wherever $\hat{\rho}$ is undefined. The last three columns concern the pooled estimator a pooled pipeline reports: the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$; its prediction, the finite-design expectation $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ of Proposition A.21 at each question's estimated moments minus the same moments' unbiased estimate of succ@k, which is $\widehat{C}_{\text{run}}$; and the fraction $\binom{R}{k} n^k \,/\, \binom{nR}{k}$ of pooled k-subsets that draw from k distinct runs, and it is the design parameter that separates the two references. Panel (a) reports the within-run correlation across the questions on which it is defined; 6 of the 200 questions drawn are excluded because 32 admissible prefixes could not be drawn for them, a prefix being admissible when it neither contains the final answer nor ends the generation, and the question stratification used in Table S7 is fixed by an independent pilot sample rather than by these data. The Arm B block is the same design carried to a harder corpus answered by a stronger model (Arm B): 64 competition mathematics questions on which an independent pilot sample scored between 0.05 and 0.95, answered by a 7B instruction-tuned model at the same $R = 32$ prefixes and $n = 32$ continuations, with a token limit long enough to hold the truncation rate at 0.0066; no question had to be dropped and 2 prefix draws in all were discarded. It arrives at the same middle regime from the opposite side, by making the questions harder rather than the model weaker, and its within-run correlation is an order of magnitude smaller (median 0.075 against 0.663), which is why the gap it measures is smaller while the same ordering holds. The Arm C block removes the admissibility condition altogether: same corpus, same model, same R and n as the Arm A block, but a run is now the first 48 new tokens of the generation whatever they turn out to be, so a prefix that already states the final answer and a prefix on which the generation has already stopped are both kept rather than redrawn; 0 draws were discarded and all 200 questions were analysed. Of its 6,400 prefixes, 0.837 would have passed the condition of the Arm A block; the other 0.163 either end the generation inside the prefix (0.108) or contain an extractable answer without ending it (0.055), the classes of Table S8(c). The two panels are estimates of two different generation processes rather than two estimates of one, so the blocks are read side by side and neither corrects the other; the within-run correlation is the higher of the two (median 0.750 against 0.663), which the kept prefixes that end the generation or already hold the answer account for. Intervals are resampled over questions. The $k = 1$ row is the negative control: no dependence term exists there, both cells carry a dash, and the measured gap's interval contains zero, on Arm B only just: its lower end is −0.0002.

Arm A: GSM8K, 1.5B model

(a) Design and within-run dependence

median $\hat{\rho}$	0.663
quartiles (Q_1, Q_3)	(0.386, 0.880)
questions on which $\hat{\rho}$ is defined	192
questions on which $\hat{\rho}$ is undefined	2
questions analysed, of those drawn	194/200
prefix draws rejected and redrawn	0.521

Excluded: 6 *questions with fewer than* 32 *admissible prefixes after* 12 *oversampled rounds.*

(b) Estimand gap by budget

k	$\overline{Y}$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}$	95% CI	second-order ref. (6)	$\widehat{C}_{\text{pooled}} - \overline{Y}$	predicted (10)	coll. free
1	0.5370	0.5368	−0.0002	[−0.0050, +0.0045]	—	−0.0002	—	1.000
2	0.5807	0.6783	+0.0976	[+0.0866, +0.1094]	0.0946	+0.0947	0.0911	0.970
4	0.6187	0.7865	**+0.1678**	[+0.1480, +0.1865]	0.1405	+0.1635	0.1602	0.828
8	0.6534	0.8640	+0.2107	[+0.1872, +0.2350]	0.1695	+0.2053	0.2026	0.396
16	0.6869	0.9190	+0.2322	[+0.2048, +0.2607]	0.1916	+0.2258	0.2232	0.012

continued on the next page

Table 19, continued.

k	$\overline{Y}$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}$	95% CI	second-order ref. (6)	$\widehat{C}_{\text{pooled}} - \overline{Y}$	predicted (10)	coll. free
32	0.7160	0.9570	+0.2410	[+0.2121, +0.2710]	0.2025	+0.2331	0.2261	< 0.001

Arm B: MATH-500, 7B model

(a) Design and within-run dependence

median $\hat{\rho}$	0.075
quartiles (Q_1, Q_3)	(0.021, 0.211)
questions on which $\hat{\rho}$ is defined	64
questions on which $\hat{\rho}$ is undefined	0
questions analysed, of those drawn	64/64
prefix draws rejected and redrawn	0.001

No question was excluded: 32 *admissible prefixes were collected for every question drawn, and* 2 *draws in all were discarded.*

(b) Estimand gap by budget

1	0.6855	0.7013	+0.0158	[−0.0002, +0.0336]	—	+0.0158	—	1.000
2	0.8081	0.8337	+0.0256	[+0.0112, +0.0407]	0.0264	+0.0248	0.0250	0.970
4	0.8765	0.9124	**+0.0360**	[+0.0174, +0.0581]	0.0299	+0.0351	0.0368	0.828
8	0.9180	0.9546	+0.0366	[+0.0184, +0.0571]	0.0190	+0.0361	0.0369	0.396
16	0.9463	0.9787	+0.0324	[+0.0150, +0.0533]	0.0106	+0.0321	0.0329	0.012
32	0.9639	0.9926	+0.0287	[+0.0135, +0.0486]	0.0064	+0.0285	0.0281	< 0.001

Arm C: GSM8K, 1.5B model, no prefix admissibility condition

(a) Design and within-run dependence

median $\hat{\rho}$	0.750
quartiles (Q_1, Q_3)	(0.392, 0.900)
questions on which $\hat{\rho}$ is defined	189
questions on which $\hat{\rho}$ is undefined	11
questions analysed, of those drawn	200/200
prefix draws rejected and redrawn	0.000

No question was excluded and no draw was discarded: this arm keeps a prefix whatever it contains, so the 32 *prefixes of every question drawn are simply its first* 32 *draws.*

(b) Estimand gap by budget

1	0.5333	0.5341	+0.0008	[−0.0042, +0.0060]	—	+0.0008	—	1.000
2	0.5745	0.6764	+0.1018	[+0.0904, +0.1131]	0.0994	+0.0988	0.0957	0.970
4	0.6073	0.7888	**+0.1814**	[+0.1628, +0.2008]	0.1555	+0.1766	0.1713	0.828
8	0.6388	0.8681	+0.2294	[+0.2031, +0.2541]	0.1848	+0.2235	0.2186	0.396
16	0.6659	0.9218	+0.2558	[+0.2276, +0.2841]	0.1918	+0.2496	0.2431	0.012
32	0.6941	0.9537	+0.2596	[+0.2295, +0.2913]	0.1561	+0.2544	0.2486	< 0.001

Table 20: Away from saturation the attempt-level interval is narrow, centred on the pooled number, and rarely contains the reference; on Arm A at $k = 16$ between a fifth and about a half of its realisations have collapsed to a point, depending on the tolerance (below). Attempt-level versus run-level intervals on Arms A, B and C, at $\delta = 0.05$ and at two budgets. The protocol is that of Table 10 with twice the runs on each side: each question's 32 prefixes are split into a 16-prefix estimation half and a 16-prefix held-out half, 50 times per question, and the held-out half supplies the held-out reference, a run-level estimate of succ@k from runs the interval did not see. Columns: interval; containment of the held-out reference; mean width; zero width, the fraction of intervals whose two endpoints coincide. All six rows of a block face the same reference and the same estimation half. The reference is itself a 16-prefix estimate; Table 33 estimates, on fitted laws, how far containment of it differs from coverage of the exact succ@k. The zero-width count depends strongly on the tolerance, because near a pooled estimate of one the endpoints crowd together without meeting: relaxing "equal" to "closer than" 10^{-12}, 10^{-9} and 10^{-6} takes the Arm A attempt-level fraction at $k = 16$ from 20.6% to 28.3%, 35.6% and 48.1%, and at $k = 4$ from 6.9% to 12.1% at the loosest of the three, while the run-level runCS and Clopper–Pearson rows stay at zero under every tolerance. The last three rows of each block are further run-level intervals on the estimation half's run indicators: the run-level percentile bootstrap; the same bootstrap with the $(R' - 1)/R'$ plug-in variance factor removed, R' being the number of runs in the estimation half; and the run-level Wald interval on the unbiased variance. When all or none of the R' runs succeed the resampled proportion is constant and these intervals are a single point, so their zero-width entries are a property of the construction, not of the tolerance. At $k = 4$ on Arm A they contain the reference 0.841, 0.841 and 0.781 of the time, against 0.892 for run-level runCS and 0.838 for run-level Clopper–Pearson, and a fraction 0.171 of their realisations is a single point; that failure mode is counted, not repaired. Arm B carries a within-run correlation an order of magnitude smaller and Arm C imposes no admissibility condition on the prefixes, so the pattern is not a property of one corpus, one strength of dependence, or one prefix condition.

Interval	Containment	Mean width	Zero width
Arm A: GSM8K, 1.5B model			
budget $k = 4$, 9,700 intervals, 50 splits per question			
attempt-level bootstrap	0.146	0.052	6.9%
run-level runCS	0.892	0.502	0%
run-level Clopper–Pearson	0.838	0.386	0%
run-level bootstrap	0.841	0.319	17.1%
run-level bootstrap, factor removed	0.841	0.328	17.1%
run-level Wald, unbiased variance	0.781	0.318	17.1%
budget $k = 16$, 9,700 intervals, 50 splits per question			
attempt-level bootstrap	0.188	0.044	20.6%
run-level runCS	0.886	0.499	0%
run-level Clopper–Pearson	0.838	0.383	0%
run-level bootstrap	0.838	0.311	19.0%
run-level bootstrap, factor removed	0.838	0.320	19.0%
run-level Wald, unbiased variance	0.789	0.312	19.0%
Arm B: MATH-500, 7B model			
budget $k = 4$, 3,200 intervals, 50 splits per question			
attempt-level bootstrap	0.208	0.029	4.8%
run-level runCS	0.922	0.374	0%
run-level Clopper–Pearson	0.902	0.288	0%
run-level bootstrap	0.900	0.147	59.4%
run-level bootstrap, factor removed	0.900	0.151	59.4%
run-level Wald, unbiased variance	0.875	0.147	59.4%
budget $k = 16$, 3,200 intervals, 50 splits per question			
attempt-level bootstrap	0.463	0.017	27.8%
run-level runCS	0.962	0.327	0%
run-level Clopper–Pearson	0.947	0.252	0%
run-level bootstrap	0.946	0.081	78.2%
run-level bootstrap, factor removed	0.946	0.083	78.2%
run-level Wald, unbiased variance	0.938	0.081	78.2%

continued on the next page

Table 20, continued.

Interval	Containment	Mean width	Zero width
Arm C: GSM8K, 1.5B model, no prefix admissibility condition			
budget $k = 4$, 10,000 intervals, 50 splits per question			
attempt-level bootstrap	0.190	0.053	11.3%
run-level runCS	0.905	0.504	0%
run-level Clopper–Pearson	0.858	0.387	0%
run-level bootstrap	0.857	0.320	19.5%
run-level bootstrap, factor removed	0.857	0.329	19.5%
run-level Wald, unbiased variance	0.807	0.318	19.5%
budget $k = 16$, 10,000 intervals, 50 splits per question			
attempt-level bootstrap	0.202	0.041	22.4%
run-level runCS	0.894	0.508	0%
run-level Clopper–Pearson	0.841	0.390	0%
run-level bootstrap	0.839	0.321	19.7%
run-level bootstrap, factor removed	0.839	0.330	19.7%
run-level Wald, unbiased variance	0.800	0.322	19.7%

Table 21: Where the shared segment is cut is part of the estimand, not a tuning knob. The same questions, model, prompt, verifier and sampling parameters as Arm A (Table 19), with the shared prefix cut at $P \in \{12, 24, 48, 96, 192\}$ new tokens and the continuation budget the mechanical remainder, $512 - P$ tokens. Each cut defines a different generation process, so the rows are five estimands placed side by side rather than five measurements of one: no curve is fitted, no test is run, and no monotonicity is asserted. The question list is the same 50 for every row, chosen by a mechanical rule. Panel (a) shows why the rows do not share a denominator: a prefix is admissible only if it neither contains the final answer nor ends the generation, and the later the cut, the more often a draw fails that condition, so questions for which 32 admissible prefixes could not be collected in twelve rounds of oversampling are dropped — 16 of them at $P = 96$ and 43 at $P = 192$. The discarded-draw and truncation columns are the two mechanisms behind that, and truncation is scored as failure throughout, so the shrinking continuation budget is a confound of the later cuts and not a separable effect. Panel (b) reports the gap on each row's own question set, with $\widehat{U}_k$ unbiased for pass@k^{iid} at every budget since $R = 32 \geq k$, alongside the finite-design ratio of Proposition A.21 over the budget grid, in the convention of Table 19: the observed $\widehat{C}_{\text{pooled}} - \overline{Y}$ over the predicted $\mathbb{E}[\widehat{C}_{\text{pooled}}] - \widehat{C}_{\text{run}}$, both averaged over questions, and the $k = 4$ containment of the held-out reference by the attempt-level bootstrap and by run-level RUNCS. Panel (c) repeats the gap on the 7 questions that no row excluded, which is the only common-denominator comparison the design admits; that denominator is small and its intervals are correspondingly wide, and it is reported because a denominator that changes with P is exactly the confound panel (b) cannot remove. Intervals are resampled over questions. †The $P = 48$ row generates nothing: at that cut Arm A already holds the data, and the row is Arm A's raw continuations restricted to the same question list and reanalysed, so it is drawn under a different sampling seed from the other four and none of the five rows is paired with any other.

(a) Design, within-run dependence, and what the admissibility condition removes

P	questions	excluded	median $\hat{\rho}$ (Q_1, Q_3)	$\hat{\mu}_1$	discarded draws	truncated
12	50/50	0	0.1915 (0.0552, 0.3834)	0.4806	0.0006	0.0049
24	50/50	0	0.3228 (0.1663, 0.5654)	0.4998	0.0025	0.0041
48†	50/50	0	0.6577 (0.4285, 0.9022)	0.4876	0.3456	0.0046
96	34/50	16	0.9371 (0.8102, 0.9629)	0.4403	0.9292	0.0105
192	7/50	43	0.8163 (0.7353, 0.8904)	0.4106	0.9935	0.0858

(b) The gap, each arm on its own question set

P	$\widehat{U}_4 - \overline{Y}$	95% CI	$\widehat{U}_{32} - \overline{Y}$	95% CI	(10) ratio range	$k = 4$ cont. attempt / run

continued on the next page

Table 21, continued.

12	+0.0743	[+0.0504, +0.0997]	+0.0700	[+0.0429, +0.1008]	1.02–1.14	0.238 / 0.895
24	+0.1187	[+0.0843, +0.1519]	+0.1348	[+0.0896, +0.1845]	1.01–1.12	0.214 / 0.910
48†	+0.1768	[+0.1362, +0.2154]	+0.2679	[+0.2061, +0.3327]	1.01–1.08	0.145 / 0.881
96	+0.2519	[+0.1954, +0.3083]	+0.3550	[+0.2674, +0.4510]	1.03–1.05	0.167 / 0.915
192	+0.2413	[+0.1675, +0.3151]	+0.5000	[+0.2768, +0.7009]	1.04–1.10	0.111 / 0.874

(c) The same quantities on the 7 questions no arm excluded (common denominator)

P	questions	median $\hat{\rho}$ (Q_1, Q_3)	$\hat{\mu}_1$	$\widehat{U}_4 - \overline{Y}$	95% CI	$\widehat{U}_{32} - \overline{Y}$
12	7	0.1616 (0.0986, 0.1915)	0.2977	+0.0751	[+0.0216, +0.1314]	+0.1212
24	7	0.1734 (0.1599, 0.3191)	0.3551	+0.1158	[+0.0443, +0.1902]	+0.1115
48†	7	0.4169 (0.1364, 0.5843)	0.2995	+0.1746	[+0.0692, +0.2836]	+0.2462
96	7	0.6567 (0.3799, 0.7093)	0.3100	+0.1799	[+0.1229, +0.2316]	+0.3628
192	7	0.8163 (0.7353, 0.8904)	0.4106	+0.2413	[+0.1663, +0.3177]	+0.5000

B.6 A SELF-DEBUGGING LOOP

This section is the complete version of the self-debugging loop of Section 6.4 of the main text.

Table 22: Against a fresh 128-run reference on Arm A the attempt-level bootstrap contains the reference 0.104 of the time, the run-level Clopper–Pearson interval 0.952 and run-level runCS 0.990. The Arm A intervals of Table 20 scored against three references at $k = 4$; the splits, the estimation halves and the six intervals are those of that table, and only the reference changes. The held-out column is the usual held-out reference, a 16-run average. The fresh column replaces it with a 128-run average drawn afresh on the same 194 questions, whose per-question standard error averages 0.0315 (binomial) against a split-to-split s.d. of 0.0628 for the 16-run held-out reference (binomial s.e. 0.0866). Both are containment figures. The fitted-law column is coverage of the exact succ@k of laws fitted to each question, so its denominator is a set of simulated questions, not the same real ones, and the last three intervals are not scored there. Zero width is the fraction of intervals whose two endpoints coincide.

	held-out	fresh	fitted law	mean width	zero width
attempt-level bootstrap	0.146	0.104	0.135	0.052	0.069
run-level runCS	0.892	0.990	0.999	0.502	0.000
run-level Clopper–Pearson	0.838	0.952	0.990	0.386	0.000
run-level bootstrap	0.841	0.789	n/a	0.319	0.171
run-level bootstrap, factor removed	0.841	0.790	n/a	0.328	0.171
run-level Wald, unbiased variance	0.781	0.783	n/a	0.318	0.171

Table 23: Both allocations of the same per-question budget give the same gap within its interval at every budget. The gap on the same questions under two allocations of the same per-question budget. The left column draws $R = 128$ prefixes with $n = 8$ continuations each, the right column is Arm A with $R = 32$ and $n = 32$; both spend 1024 continuations per question. The two are not repeated measurements of one estimator, so the reading is whether the gap has the same sign and the same order of magnitude, not whether the numbers agree. Budgets $k > n$ have no unbiased run-level estimator at $n = 8$ and are omitted by design. Intervals are percentile bootstrap over questions.

	$R = 128$, $n = 8$ (fresh)	$R = 32$, $n = 32$ (Arm A)
(a) dependence strength		
$\hat{\rho}$, median	0.718	0.663
$\hat{\rho}$, first quartile	0.489	0.386
$\hat{\rho}$, third quartile	0.881	0.880
first-attempt rate $\bar{Y}_1$	0.5358 [0.4869, 0.5819]	0.5370 [0.4892, 0.5842]
(b) the estimand gap and the pooled excess		
$\hat{U}_1 - \bar{Y}$	+0.0004 [−0.0023, +0.0031]	−0.0002 [−0.0052, +0.0049]
$\hat{U}_2 - \bar{Y}$	+0.0992 [+0.0893, +0.1090]	+0.0976 [+0.0867, +0.1089]
$\hat{U}_4 - \bar{Y}$	+0.1704 [+0.1535, +0.1891]	+0.1678 [+0.1494, +0.1869]
$\hat{U}_8 - \bar{Y}$	+0.2161 [+0.1912, +0.2401]	+0.2107 [+0.1873, +0.2355]
$\hat{C}_{\text{pooled}} - \bar{Y}$ at $k = 1$	+0.0004 [−0.0022, +0.0032]	−0.0002 [−0.0050, +0.0046]
$\hat{C}_{\text{pooled}} - \bar{Y}$ at $k = 2$	+0.0986 [+0.0889, +0.1085]	+0.0947 [+0.0835, +0.1051]
$\hat{C}_{\text{pooled}} - \bar{Y}$ at $k = 4$	+0.1695 [+0.1524, +0.1868]	+0.1635 [+0.1450, +0.1812]
$\hat{C}_{\text{pooled}} - \bar{Y}$ at $k = 8$	+0.2148 [+0.1915, +0.2386]	+0.2053 [+0.1827, +0.2292]

Table 24: On the 1.5B model, four independent fresh attempts succeed more often than four rounds of the self-debugging loop, with an interval above zero from $k = 2$; on the 7B model the gap is indistinguishable from zero. The estimand gap measured on a program rather than on a sampler: the self-debugging loop on the 7B model (first block). Each run is one complete execution of a Reflexion-style self-debugging loop on a Python synthesis benchmark: the model first writes its own internal tests, then submits an implementation, and while its own tests fail and the budget allows it reflects on the failure and resubmits, stopping as soon as its internal verifier passes. The hidden tests are never shown to the model and never enter a prompt; they are evaluated in a network-isolated sandbox. 164 questions carry $R = 16$ independent runs each, and a run may spend up to $k_{\max} = 4$ attempts. $\overline{Y}$ estimates succ@k for this program without bias under its own stopping rule (Proposition 3.4); $\widehat{U}_k$ is the cross-run unbiased estimate of pass@k^{iid} built from the *first* attempt of each run, which is the only attempt of a run with an i.i.d. counterfactual, because every later attempt is conditioned on the internal verifier having rejected the previous one. Their difference is therefore an unbiased estimate of pass@k^{iid} − succ@k on the same denominator. The last column keeps the pooled estimator a pooled pipeline reports; here it is descriptive only, since the pooled attempts of a run are neither exchangeable nor identically distributed and Proposition A.21 does not apply. Panel (a) records what the loop actually did: how many attempts runs spent, how often the internal verifier accepted a wrong program or rejected a right one, and how often a rejected attempt was repaired by the next one. Intervals are percentile bootstrap over questions. The $k = 1$ row is zero by construction: at a budget of one attempt the two estimators are the same statistic. The second block repeats the identical protocol on a smaller model of the same family, 1.5B, whose first-attempt success rate is $\hat{\mu}_1 = 0.5271$ rather than 0.8190. The attempts are ordered, so (6) does not predict this loop's gap; these are descriptive comparisons of leading-budget success with an independent-first-attempt counterfactual.

k	$\overline{Y}$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}$	95% CI	$\widehat{C}_{\text{pooled}} - \overline{Y}$
Self-debugging loop, 7B instruction-tuned model					
(a) Design, stopping and the internal verifier					
questions analysed, of those drawn					164/164
runs per question R, attempt budget $k_{\max}$					16×4
runs, attempts generated					2,624, 5,858
$\hat{\mu}_1$, first attempt, mean over questions					0.8190 [0.7687, 0.8662]
questions with $\hat{\mu}_1 > 0.9$					113/164
runs stopping after 1, 2, 3, 4 attempts					1,471, 98, 29, 1,026
mean attempts per run $\bar{\tau}$					2.232
runs whose internal test set is empty					0%
internal verifier, false accept $\mathbb{P}(X = 0 \mid V = 1)$					0.0476 77/1,618
internal verifier, false reject $\mathbb{P}(X = 1 \mid V = 0)$					0.5462 2,316/4,240
repair rate $\mathbb{P}(X_{i+1} = 1 \mid X_i = 0,\ V_i = 0)$					0.2000 285/1,425
sandbox executions, of which timed out					11,890, 1
prompts scanned, windows hit, unassigned					8,482, 652, 1
Every prompt is a deterministic function of four inputs: the fixed instruction template, the public problem statement, the model's own previous implementation and the failure summary of the model's own internal tests. The hidden test string is not one of them. The scan nevertheless finds 652 *of the* 8,482 *prompts carrying a window of* 40 *characters that also occurs in the hidden test string, because the official tests reuse the worked examples of the public docstring verbatim and so do the model's own asserts. Of the* 14,286 *such windows,* 14,285 *lie inside one of the four inputs; the remaining* 1 *straddles the join between two of them and so lies inside neither, which is what the assignment rule reports rather than any occurrence of test text in a prompt.*					
(b) Estimand gap by budget					
1	0.8190	0.8190	0.0000	[0.0000, 0.0000]	−0.0586
2	0.8605	0.8666	+0.0061	[−0.0073, +0.0193]	−0.0086
3	0.8735	0.8836	+0.0101	[−0.0071, +0.0264]	+0.0128
4	0.8826	0.8918	**+0.0092**	[−0.0116, +0.0271]	+0.0205
Self-debugging loop, 1.5B instruction-tuned model					
(a) Design, stopping and the internal verifier					
questions analysed, of those drawn					164/164
runs per question R, attempt budget $k_{\max}$					16×4
runs, attempts generated					2,624, 7,738
$\hat{\mu}_1$, first attempt, mean over questions					0.5271 [0.4684, 0.5850]

continued on the next page

Table 24, continued.

k	$\overline{Y}$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}$	95% CI	$\widehat{C}_{\text{pooled}} - \overline{Y}$
questions with $\hat{\mu}_1 > 0.9$					50/164
runs stopping after 1, 2, 3, 4 attempts					864, 58, 50, 1,652
mean attempts per run $\bar{\tau}$					2.949
runs whose internal test set is empty					0%
internal verifier, false accept $\mathbb{P}(X = 0 \mid V = 1)$					0.1015 101/995
internal verifier, false reject $\mathbb{P}(X = 1 \mid V = 0)$					0.2751 1,855/6,743
repair rate $\mathbb{P}(X_{i+1} = 1 \mid X_i = 0,\ V_i = 0)$					0.0753 274/3,640
sandbox executions, of which timed out					15,650, 1
prompts scanned, windows hit, unassigned					10,362, 987, 2

Every prompt is a deterministic function of four inputs: the fixed instruction template, the public problem statement, the model's own previous implementation and the failure summary of the model's own internal tests. The hidden test string is not one of them. The scan nevertheless finds 987 *of the* 10,362 *prompts carrying a window of* 40 *characters that also occurs in the hidden test string, because the official tests reuse the worked examples of the public docstring verbatim and so do the model's own asserts. Of the* 24,647 *such windows,* 24,645 *lie inside one of the four inputs; the remaining* 2 *straddle the join between two of them and so lie inside neither, which is what the assignment rule reports rather than any occurrence of test text in a prompt.*

(b) Estimand gap by budget

1	0.5271	0.5271	0.0000	[0.0000, 0.0000]	−0.0995
2	0.5625	0.6322	+0.0697	[+0.0520, +0.0876]	−0.0005
3	0.5869	0.6876	+0.1007	[+0.0750, +0.1241]	+0.0471
4	0.5983	0.7236	**+0.1253**	[+0.0964, +0.1526]	+0.0819

Table 25: On both models and at both budgets, run-level intervals contain the held-out reference more often than the attempt-level bootstrap. Attempt-level versus run-level intervals when the unit is a run of the self-debugging loop, on the 7B and 1.5B models, at $\delta = 0.05$ and at two budgets. The protocol is that of Table 10 with the run in place of the shared prefix: each question's 16 runs are split into an estimation half and a held-out half, 50 times per question, and the held-out half supplies the held-out reference, a run-level estimate of succ@k from runs the interval did not see. Columns: interval; containment of the held-out reference; mean width; zero width, the fraction of intervals whose two endpoints coincide. The attempt-level interval pools attempts of different rank inside a run, which is exactly the population the loop's stopping rule distorts. The last three rows of each block are the further run-level intervals of Table 20; at $k = 4$ on the 7B model they contain the reference 0.919, 0.919 and 0.912 of the time against 0.925 for run-level runCS, with a fraction 0.822 of their realisations a single point. These designs have no fitted-law calibration, so the correction of Table 32 is not transferred here; the absolute levels are read against each other and not against a nominal 0.95.

Interval	Containment	Mean width	Zero width
Self-debugging loop, 7B model			
budget $k = 2$, 8,200 intervals, 50 splits per question			
attempt-level bootstrap	0.785	0.125	51.6%
run-level runCS	0.944	0.491	0%
run-level Clopper–Pearson	0.940	0.413	0%
run-level bootstrap	0.935	0.106	80.0%
run-level bootstrap, factor removed	0.935	0.111	80.0%
run-level Wald, unbiased variance	0.926	0.103	80.0%
budget $k = 4$, 8,200 intervals, 50 splits per question			
attempt-level bootstrap	0.785	0.078	55.8%
run-level runCS	0.925	0.482	0%
run-level Clopper–Pearson	0.920	0.406	0%
run-level bootstrap	0.919	0.084	82.2%
run-level bootstrap, factor removed	0.919	0.088	82.2%
run-level Wald, unbiased variance	0.912	0.087	82.2%

continued on the next page

Table 25, continued.

Interval	Containment	Mean width	Zero width
Self-debugging loop, 1.5B model			
budget $k = 2$, 8,200 intervals, 50 splits per question			
attempt-level bootstrap	0.692	0.256	25.2%
run-level runCS	0.880	0.591	0%
run-level Clopper–Pearson	0.867	0.498	0%
run-level bootstrap	0.869	0.307	47.3%
run-level bootstrap, factor removed	0.869	0.323	47.3%
run-level Wald, unbiased variance	0.827	0.303	47.3%
budget $k = 4$, 8,200 intervals, 50 splits per question			
attempt-level bootstrap	0.670	0.248	27.4%
run-level runCS	0.887	0.589	0%
run-level Clopper–Pearson	0.873	0.497	0%
run-level bootstrap	0.860	0.295	47.2%
run-level bootstrap, factor removed	0.860	0.310	47.2%
run-level Wald, unbiased variance	0.814	0.300	47.2%

The loop's transition law never reads $k_{\max}$, so Assumption A.2 holds and the $k < 4$ rows of Table 24 are G_k. Two things about this procedure decide what the two statistics mean. The shared state of a run is the internal test set drawn once at its start, so attempts within a run are dependent through it and, in addition, ordered: attempt $i + 1$ exists only because the verifier rejected attempt i. The only attempt with an i.i.d. counterfactual is therefore the first, and $\widehat{U}_k$ is built from the R first attempts, which makes it unbiased for pass@k^{iid}, the success probability of k independent fresh attempts. The run-level indicator is unbiased for succ@k under the loop's own stopping rule (Proposition 3.4), which here is the truncated procedure of Proposition A.17: a run whose verifier accepts a wrong program stops on a failure, and that is a property of the procedure, not of the measurement. The pooled estimator is kept for reference only; its attempts are neither exchangeable nor identically distributed, and Proposition A.21 does not apply. A scan for forty-character windows shared with the hidden tests places every such window inside one of a prompt's four inputs except one at 7B and two at 1.5B, which straddle the join between two inputs; none is test text in a prompt (Table 24(a)).

On the 7B model the first attempt succeeds 0.8190 of the time, 113 of the 164 problems sit above 0.9, and 56.1% of runs stop after one attempt; the gap $\widehat{U}_4 - \overline{Y}$ is +0.0092 with an interval from −0.0116 to +0.0271, indistinguishable from zero; the loop's attempts are ordered, so (6) is not a reference for it and no mechanism is attached to the size (§6.4). On the 1.5B model the first attempt succeeds 0.5271 of the time and the loop has work to do — 63.0% of runs spend all four attempts — and the gap is +0.1253 (interval +0.0964 to +0.1526) at $k = 4$ and +0.0697 at $k = 2$: four independent fresh attempts succeed more often than four rounds of the loop's own revision, by 0.1253 in success probability, on the same problems and the same model. The loop is not idle; its verifier accepts a wrong program 0.1015 of the time it accepts, rejects a right one 0.2751 of the time it rejects, and a rejected attempt is repaired by the next one 0.0753 of the time. What the number says is that the state a run carries — its own tests, its own draft — makes its later attempts a poorer draw than a fresh start, and that the independent-sample pass@4^{iid} exceeds the loop's own succ@4 by an estimated 0.1253. The containment comparison (Table 25) is the one seen throughout: with only four attempts per run the attempt-level interval is not as narrow as on Arms A–C, but it contains the held-out reference 0.670 of the time at $k = 4$ against 0.887 for runCS and 0.873 for the run-level Clopper–Pearson interval, and 27.4% of its realisations have zero width.

B.7 A tree-search scaffold

This section is the complete version of the tree-search scaffold of Section 6.4 of the main text.

Table 26: The tree-search scaffold shows a small positive estimand gap at $k = 2$ and $k = 4$, with an interval above zero under every variation. A tree-search scaffold with no constructed dependence, on the same 194 arithmetic questions as Arm A and the same 7B model as the headline design. A run is one breadth-first search of depth two and branching two, scored and pruned to the best two nodes by the model's own value prompt, yielding 4 leaves; 16 fresh runs per question. No prefix is imposed, no candidate is resampled, and nothing about the reference answer enters generation or pruning. Panel (a): the run-level success probability $\overline{Y}_k$, the cross-run statistic $\widehat{U}_k$ on the first leaf of each run (the first-attempt form, unbiased for pass@k^{iid}), and their difference, beside the pooled reading $\widehat{C}_{\text{pooled}}$ and the cross-run statistic $\widehat{U}_k^{\text{all}}$ of Proposition A.8 applied to all leaves. At $k = 1$, $\overline{Y}_k$ equals the first-leaf cross-run statistic, while $\widehat{C}_{\text{pooled}}$ equals the all-leaf statistic; these two marginal targets need not coincide. Panel (b): the $k = 4$ reading under three variations — the alternative leaf ordering, the runs in which no node terminated early, and the questions on which every value call parsed (bootstrap seed differs; canonical interval in Table 50). Panel (c): containment of the two interval families at $k = 4$ over 9,700 intervals, scored against a held-out half of 8 runs and against all 16 runs. Intervals are percentile bootstrap over questions, $B = 2{,}000$; the within-run correlation has median 0.205 (quartiles 0.078 and 0.370) and the leaf-level success rate is 0.893. Early stopping inside a step is a real property of the scaffold and is reported rather than filtered: 0.000 of first-step nodes and 0.270 of second-step nodes already carry a final answer, so 0.287 of leaves repeat a completed chain, and the value prompt returns one of its three labels on 1.000 of calls.

(a) the estimand gap of the search scaffold, by budget

k	$\hat{\mu}_1$	$\hat{\rho}$	$\overline{Y}_k$	$\widehat{U}_k$	$\widehat{U}_k - \overline{Y}_k$ mean	95% CI	$\widehat{C}_{\text{pooled}}$	$\widehat{U}_k^{\text{all}}$	$\widehat{C}_{\text{pooled}} - \widehat{U}_k^{\text{all}}$
1	0.8934	0.2045	0.8882	0.8882	0.0000	[+0.0000, +0.0000]	0.8934	0.8934	0.0000
2	0.8934	0.2045	0.9040	0.9231	+0.0191	[+0.0110, +0.0284]	0.9302	0.9308	−0.0006
4	0.8934	0.2045	0.9346	0.9480	+0.0134	[+0.0016, +0.0257]	0.9559	0.9569	−0.0010

(b) the $k = 4$ reading under three variations

variation	questions	$\hat{\mu}_1$	$\hat{\rho}$	$\widehat{U}_4 - \overline{Y}_4$	
completion order, all runs (main estimate)	194	0.8934	0.2045	+0.0134	[+0.0016, +0.0257]
generation order, all runs	194	0.8934	0.2045	+0.0190	[+0.0083, +0.0310]
runs with no early-terminating node	140	0.8729	0.2045	+0.0173	[+0.0028, +0.0332]
questions with every value parsed	194	0.8934	0.2045	+0.0134	[+0.0014, +0.0258]

(c) containment at $k = 4$, and interval width

interval	held-out half	all runs	mean width	median width
attempt-level bootstrap	0.913	0.917	0.041	0.000
run-level runCS	0.984	1.000	0.473	0.439
run-level Clopper–Pearson	0.976	1.000	0.398	0.369

The value prompt is answered greedily and never fails to parse: it returns *sure* on 0.2590 of calls, *likely* on 0.6674 and *impossible* on 0.0735, with 0.0000 unparsed, so the prompt is doing work rather than collapsing onto one label. Ties are broken by a per-run random permutation, and the two survivors share a parent in 0.4046 of runs. A node that emits its end-of-sequence token inside its own segment has already written a final answer and is not expanded further, which happens for 0.0005 of first-step nodes and 0.2705 of second-step ones and makes 0.2867 of leaves a repeat of a completed chain.

Four things are worth reading against the body. First, the correlation is undefined on 141 of the 194 questions, whose runs are unanimous: there $\overline{Y}$, $\widehat{U}_k$ and $\widehat{C}_{\text{pooled}}$ coincide, the contribution to the gap is exactly zero, and the questions stay in every denominator, so the reported gap is carried by the remaining questions and diluted across all of them. Second, the second-order reference of (6) overstates the $k = 4$ gap by about a factor of five, because the four leaves of a tree are not exchangeable — two of the six pairs share a deeper prefix than the other four — and the latent-variable expansion assumes conditionally i.i.d. attempts; the direction is right and the magnitude is not, which is why this row of Table 7 carries no finite-design ratio. Third, about a quarter of first-step nodes coincide byte for byte with a first-step node of another run of the same question. That is not a violation of Assumption 5.1 — independent draws may coincide, and no state passes between runs through the inference cache — but it says the sixteen runs of a question explore fewer than sixteen distinct first steps at this model and temperature. Fourth, the success rate rises along the completion order, 0.8882, 0.8921, 0.8956 and 0.8976: the leaf that finishes first is slightly the worst *on the corpus-wide marginals*. That ordering leaves open which of the four position-specific counterfactuals is the most conservative: ordering four global means does not order the equally weighted average over questions of the per-question independent-run counterfactuals, which is what the main estimate reports: g_k is not linear, so a position with the lower global mean can have the higher mean counterfactual. Settling the direction would need the four matched per-question readings with their paired uncertainty, which are not computed here; the direction is left open. Table 18 is what that calculation looks like on the reasoning tree of §6.2, whose position effect has the opposite sign; both are reported.

Panel (b) of Table 26 collects three variations: renumbering the leaves by generation order rather than completion order gives +0.0190; the runs containing no early-terminating node, which leave 140 questions, give +0.0173, so the main estimate does not rest on the early-stopping convention; and restricting to questions on which every value call parsed leaves the main estimate unchanged. All four intervals exclude zero. Panel (c) reads the two interval families over 9,700 intervals at $k = 4$: the attempt-level bootstrap contains the held-out reference 0.913 of the time at mean width 0.041, against 0.984 at 0.473 for runCS and 0.976 at 0.398 for the run-level Clopper–Pearson interval. The contrast is narrower than on the shared-prefix designs for a reason the widths make visible: at this success rate most of the pooled attempt-level intervals are degenerate, and a zero-width interval on a question whose target is 1 is scored as a hit.

C Simulation and calibration

This appendix checks the predictions against a known truth: the controlled-truth simulation (§C.1), the calibration of the noisy held-out references and the cost of the designs (§C.2), and the decision experiment (§C.3). Every number is produced by the same artefacts as the corresponding table in the body; the full grids are in the supplementary report (Supplement §§S3–S4).

C.1 Controlled truth: simulation

On a generator whose succ@k and pass@k^{iid} are both available in closed form, every prediction of §4 and §5 becomes a numerical assertion that either holds cell by cell or does not.

Tables 27 and 28 report the grid. Across the generator grid the pooled classical estimator equals the closed-form succ@k cell by cell and sits below pass@k^{iid} by exactly the finite expansion of Corollary A.5; the largest bias over the two-point family reaches the $R = 1$ minimax bound of Theorem A.6 to within a factor of two (Table S16), so the bound is not loose in practice; and as the number of attempts per run grows, finite-sample coverage patterns are compared with the limsup upper bound of Proposition A.22 while runCS stays above its nominal level in every cell. Table 29 repeats the check on a generator whose attempts are dependent but not exchangeable: there the pooled classical estimator targets the subset average of Proposition 3.3, not succ@k.

Table 27: The pooled classical estimator equals the closed-form succ@k within Monte-Carlo error in every exchangeable cell and there sits below pass@k^{iid} by the exact expansion. Simulation with closed-form truth. Eight generators with a common single-attempt rate $\mu_1 = 0.4$ and four levels of latent variance, at budget $k = 4$, $R = 1$ run and $n = 64$ attempts per run, over 10,000 independent replicates per cell. Columns: generator family and latent variance; the within-run correlation ρ implied by that variance; the closed-form succ@k and pass@k^{iid}; the Monte-Carlo mean of the pooled classical estimator with its standard error; the standardised distance from the pooled estimator to succ@k; and the distance to pass@k^{iid}. The prediction under test is that the $z(\widehat{C}_{\text{pooled}} - \text{succ@}k)$ column is within Monte-Carlo error of zero in every cell while the $\widehat{C}_{\text{pooled}} - \text{pass@}k^{\text{iid}}$ column tracks minus the exact expansion (6). The first Beta row is the degenerate limit of the family, a point mass at μ_1. The last row replaces the exchangeable draw by a sequential self-correction kernel over 500,000 replicates, for which the run records no latent variance and no intraclass correlation, and estimates succ@k itself by Monte Carlo (standard error 0.00039), so no standardised distance is stored; the corresponding stored quantity is $\widehat{C}_{\text{pooled}} - \text{succ@}k = 0.0792$.

Generator	Var(θ)	ρ	succ@k	pass@k^{iid}	$\mathbb{E}[\widehat{C}_{\text{pooled}}] \pm$ SE	$z(\widehat{C}_{\text{pooled}} - \text{succ@}k)$	$\widehat{C}_{\text{pooled}} - \text{pass@}k^{\text{iid}}$
two-point	0.0000	0.0000	0.8704	0.8704	0.8708 ± 0.00054	+0.69	+0.0004
two-point	0.0100	0.0417	0.8487	0.8704	0.8481 ± 0.00108	−0.58	−0.0223
two-point	0.0400	0.1667	0.7824	0.8704	0.7823 ± 0.00205	−0.04	−0.0881
two-point	0.0900	0.3750	0.6679	0.8704	0.6700 ± 0.00333	+0.63	−0.2004
Beta	0.0000	0.0000	0.8704	0.8704	0.8704 ± 0.00054	−0.08	0.0000
Beta	0.0100	0.0417	0.8489	0.8704	0.8483 ± 0.00113	−0.55	−0.0221
Beta	0.0400	0.1667	0.7857	0.8704	0.7870 ± 0.00221	+0.60	−0.0834
Beta	0.0900	0.3750	0.6844	0.8704	0.6839 ± 0.00340	−0.16	−0.1865
Within-run non-exchangeable kernel (sequential self-correction)							
sequential	n/a	n/a	0.9182	0.9375	0.9975 ± 0.00004	n/a	+0.0600

C.2 Finite designs and the cost of a noisy target

Two further controlled-truth checks address the finite designs of §6 directly. Table S9 scores each estimator against the quantity it is unbiased for on the (R, n) pairs the evidence uses, including two laws matched to the two real corpora by first moment and intraclass correlation; the finite-design expectation (10) holds in every cell (the largest of the table's 120 Monte-Carlo z-scores is 3.40, with chance probability 0.08 at that count). Table 30 checks the same expectation out of sample on the three real arms, with the moments fitted on one half of each question's runs and the pooled number observed on the other; Table S10 scores the parametric plug-in for pass@k^{iid} against $\widehat{U}_k$ on laws fitted to the same arms; Table 31 gives an assumption-dependent sensitivity calculation for questions with undefined within-run correlation; and Table 36 carries the design comparison of §F.3. Feeding the formula unbiased moment estimates from a single design, as Table 9 does, biases the plug-in downward at $R = 16$ by an amount small against the gap, which is the direction and order of the residual seen on the shared-prefix data; with the four groups per question that the reasoning tree typically offers the same plug-in is no longer negligible against the quantity predicted, which is why Table 17 carries no finite-design column. Table 32 replays the containment protocol of Table 10 on laws fitted question by question and reports containment of the held-out reference the real protocol uses and coverage of the exact truth, which the real protocol cannot see. Table 33 does the same for the other design the evidence uses, that of §6.3: $R = 32$ runs of $n = 32$ attempts, budgets $k \in \{4, 16\}$, a sixteen-run estimation half against a sixteen-run held-out half, 194 questions, with each per-question law fixed by two moments — a Beta(a, b) on 180 questions, a two-point law on $\{0, 1\}$ on 3, and a point mass on 11. Table 35 does the same for the reasoning tree of §6.2, keeping every question's own unbalanced number of sibling groups: there the held-out half holds as few as two groups, and against the exact truth runCS covers 0.998 at $k = 4$ where the measured figure was 0.788, while the attempt-level interval moves only from 0.609 to 0.640. Table 38 asks the question one step further on, at the point where an interval is used to decide something, and is described in §5.3. Table 37 is of a different kind: it measures, on the generation runs themselves, the cost ratio that Proposition A.24 takes as given, and re-evaluates Table 5 at the measured value. Three further tables score the same designs against held-out reference: Table S13 gives every row of Table 36 its measured held-out bias and containment, the pooled interval containing the held-out reference in 0.000, 0.120 and 0.580 of splits on the three Arm A rows with $R < k$ against 0.930 to 1.000 once $R \geq k$; Table 34 puts

Table 28: As the attempts per run grow, attempt-level coverage of succ@k collapses toward zero, while run-level runCS stays above its nominal level. Coverage collapse against Proposition A.22. Design G2, $\theta \sim \text{Beta}(2, 2)$: $\mu_1 = 0.50$, $\text{Var}(\theta) = 0.05$, succ@$k = 0.8571$ and pass@$k^{\text{iid}} = 0.9375$ in closed form at budget $k = 4$, with $\delta = 0.05$, 2,000 independent replicates per cell and 2,000 bootstrap resamples inside each interval. Common random numbers are used across the n sweep — the latent draw and the first k attempts of every replicate are shared — so the two run-level columns are constant down a row block by construction, and only the attempt-level columns move with n. Columns: number of independent runs R; attempts per run n; empirical coverage and mean width of the attempt-level bootstrap interval and of the attempt-level exact binomial interval, each evaluated against succ@k; empirical coverage and mean width of run-level runCS of (17); and the limit $\mathbb{P}(L = \text{succ@}k)$ bounding the limsup in (11). The behavior shown is finite-sample attempt-level coverage relative to the last column, which is not a general equality claim; run-level coverage targets $1 - \delta$ with a width that does not move with n. The full grid, the second design and the two remaining interval constructions are in Tables S17 and S18. On this design the limit law of the attempt-level endpoint has no atoms, so the limsup upper bound is 0 at every R; the two-point design of Table S20 has atoms and a nonzero upper bound for the iid target; a plateau is not implied in general.

R	n	attempt-level bootstrap		attempt-level exact		run-level runCS		predicted
		coverage	width	coverage	width	coverage	width	limit
1	4	0.484	0.484	0.853	0.602	1.000	0.975	0.000
1	16	0.515	0.323	0.838	0.305	1.000	0.975	0.000
1	64	0.310	0.164	0.294	0.140	1.000	0.975	0.000
1	256	0.166	0.082	0.134	0.066	1.000	0.975	0.000
1	4096	0.041	0.020	0.029	0.016	1.000	0.975	0.000
2	4	0.752	0.459	0.905	0.455	1.000	0.896	0.000
2	16	0.538	0.215	0.664	0.202	1.000	0.896	0.000
2	64	0.312	0.106	0.256	0.095	1.000	0.896	0.000
2	256	0.167	0.053	0.139	0.046	1.000	0.896	0.000
2	4096	0.044	0.013	0.033	0.011	1.000	0.896	0.000
4	4	0.699	0.312	0.964	0.292	1.000	0.756	0.000
4	16	0.490	0.141	0.481	0.134	1.000	0.756	0.000
4	64	0.262	0.069	0.222	0.064	1.000	0.756	0.000
4	256	0.138	0.034	0.108	0.031	1.000	0.756	0.000
4	4096	0.038	0.009	0.031	0.008	1.000	0.756	0.000
8	4	0.595	0.191	0.821	0.193	0.998	0.607	0.000
8	16	0.333	0.093	0.286	0.092	0.998	0.607	0.000
8	64	0.149	0.046	0.115	0.044	0.998	0.607	0.000
8	256	0.069	0.023	0.055	0.022	0.998	0.607	0.000
8	4096	0.020	0.006	0.016	0.005	0.998	0.607	0.000

question-cluster bootstrap intervals on the $k = 4$ containment and zero-width figures of Tables 10, 20 and 25 (every budget: supplementary Table S14); and Table 39 runs the procedures of Table 38 on the measured runs of §6.3, with the simulated attempt-level type I alongside (all eight procedures: supplementary Table S25).

Table 29: Without exchangeability the pooled classical estimator lies above succ@k at every budget, by +0.4531 at $k = 1$ down to +0.0020 at $k = 16$. The within-run non-exchangeable generator across budgets: a sequential self-correction generator in which the success probability of attempt i rises with i, so the attempts inside a run are dependent but not exchangeable. Every quantity here is Monte Carlo over 500,000 replicates at $R = 1$, $n = 64$, except pass@k^{iid}, which is closed form. Reading: the pooled classical estimator is above succ@k at every budget, which is Proposition 3.3 — without exchangeability the estimator targets the subset average, not the leading-k probability — and the sign of pass@k^{iid} − succ@k no longer has to follow the exchangeable expansion. The last column is the analogue of the gap column of Table S15, here measured rather than closed form.

k	succ@k (MC)	SE	pass@k^{iid}	$\mathbb{E}[\widehat{C}_{\text{pooled}}]$	SE	$\widehat{C}_{\text{pooled}}$ vs succ@k	$\widehat{C}_{\text{pooled}}$ vs pass@k^{iid}	gap (MC)
1	0.4995	0.00071	0.5000	0.9527	0.00013	+0.4531	+0.4527	+0.0005
2	0.7488	0.00061	0.7500	0.9905	0.00007	+0.2417	+0.2405	+0.0012
4	0.9182	0.00039	0.9375	0.9975	0.00004	+0.0792	+0.0600	+0.0193
8	0.9832	0.00018	0.9961	0.9992	0.00002	+0.0160	+0.0031	+0.0129
16	0.9978	0.00007	1.0000	0.9998	0.00001	+0.0020	−0.0002	+0.0022

Table 30: Fitted on one half of each question's runs, the finite-design expectation of Proposition A.21 matches the pooled number observed on the other half to within 0.0113 on all three arms, with a small systematic downward plug-in bias. The expectation checked out of sample. For each question the R runs are split at random into two halves; the moments $\hat{m}_c$ are fitted on one half and the expectation is evaluated at design $(R/2,\ n)$, and the other half supplies the observed $\widehat{C}_{\text{pooled}}$ and $\overline{Y}$. Averages are taken over 50 splits and then over questions, and the interval is a percentile bootstrap over questions. The last three columns repeat the comparison with the moments fitted on the same runs that are then compared against, at design $(R,\ n)$, which is what the tables of §6 report: those two designs are different and the columns are not comparable cell by cell. The upper block compares the pooled number itself, the lower block the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$, predicted as $\mathbb{E}[\widehat{C}_{\text{pooled}}]$ minus the fitted half's $\widehat{C}_{\text{run}}$; it exists only for $k \le n$.

k	cross-fitted, design $(R/2,\ n)$					same data, design $(R,\ n)$		
	predicted	observed	difference	95% CI	ratio	predicted	observed	ratio
the pooled number itself								
Headline design: GSM8K, 7B model: $R = 16$, $n = 8$, 200 questions								
1	0.8957	0.8960	+0.0003	[−0.0009, +0.0017]	1.00×	0.8959	0.8959	1.00×
2	0.9287	0.9313	+0.0026	[+0.0011, +0.0043]	1.00×	0.9308	0.9320	1.00×
4	0.9519	0.9550	+0.0031	[+0.0013, +0.0052]	1.00×	0.9546	0.9560	1.00×
8	0.9674	0.9702	+0.0027	[+0.0010, +0.0046]	1.00×	0.9699	0.9714	1.00×
16	0.9771	0.9793	+0.0022	[+0.0008, +0.0038]	1.00×	0.9793	0.9805	1.00×
32	0.9816	0.9833	+0.0017	[+0.0006, +0.0031]	1.00×	0.9835	0.9843	1.00×
Arm A: GSM8K, 1.5B model: $R = 32$, $n = 32$, 194 questions								
1	0.5366	0.5370	+0.0004	[−0.0018, +0.0028]	1.00×	0.5368	0.5368	1.00×
2	0.6665	0.6726	+0.0061	[+0.0037, +0.0088]	1.01×	0.6724	0.6754	1.00×
4	0.7692	0.7774	+0.0082	[+0.0054, +0.0113]	1.01×	0.7777	0.7822	1.01×
8	0.8430	0.8521	+0.0091	[+0.0060, +0.0125]	1.01×	0.8532	0.8587	1.01×
16	0.8946	0.9050	+0.0104	[+0.0070, +0.0141]	1.01×	0.9062	0.9126	1.01×
32	0.9294	0.9407	+0.0113	[+0.0073, +0.0157]	1.01×	0.9421	0.9491	1.01×
Arm B: MATH-500, 7B model: $R = 32$, $n = 32$, 64 questions								
1	0.7013	0.7013	0.0000	[−0.0017, +0.0019]	1.00×	0.7013	0.7013	1.00×
2	0.8303	0.8320	+0.0017	[+0.0003, +0.0035]	1.00×	0.8320	0.8329	1.00×
4	0.9084	0.9106	+0.0022	[+0.0008, +0.0038]	1.00×	0.9105	0.9115	1.00×
8	0.9517	0.9535	+0.0017	[+0.0002, +0.0035]	1.00×	0.9533	0.9540	1.00×
16	0.9770	0.9779	+0.0009	[−0.0007, +0.0025]	1.00×	0.9778	0.9784	1.00×
32	0.9916	0.9919	+0.0003	[−0.0010, +0.0012]	1.00×	0.9920	0.9924	1.00×

continued on the next page

Table 30, continued.

k	cross-fitted, design $(R/2,\ n)$					same data, design $(R,\ n)$		
	predicted	observed	difference	95% CI	ratio	predicted	observed	ratio
the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$								
Headline design: GSM8K, 7B model: $R = 16$, $n = 8$, 200 questions								
1	0.0000	−0.0021	−0.0021	[−0.0071, +0.0033]	n/a	0.0000	−0.0023	n/a
2	0.0122	0.0112	−0.0010	[−0.0054, +0.0032]	0.92×	0.0141	0.0123	0.87×
4	0.0200	0.0240	+0.0040	[−0.0001, +0.0087]	1.20×	0.0223	0.0251	1.12×
8	0.0221	0.0243	+0.0022	[+0.0006, +0.0040]	1.10×	0.0243	0.0257	1.06×
Arm A: GSM8K, 1.5B model: $R = 32$, $n = 32$, 194 questions								
1	0.0000	−0.0004	−0.0004	[−0.0051, +0.0044]	n/a	0.0000	−0.0002	n/a
2	0.0853	0.0917	+0.0064	[+0.0020, +0.0110]	1.07×	0.0911	0.0947	1.04×
4	0.1517	0.1584	+0.0067	[+0.0023, +0.0110]	1.04×	0.1602	0.1635	1.02×
8	0.1925	0.1988	+0.0063	[+0.0015, +0.0112]	1.03×	0.2026	0.2053	1.01×
16	0.2118	0.2176	+0.0059	[+0.0007, +0.0110]	1.03×	0.2232	0.2258	1.01×
32	0.2139	0.2242	+0.0103	[+0.0063, +0.0151]	1.05×	0.2261	0.2331	1.03×
Arm B: MATH-500, 7B model: $R = 32$, $n = 32$, 64 questions								
1	0.0000	0.0167	+0.0167	[−0.0006, +0.0362]	n/a	0.0000	0.0158	n/a
2	0.0234	0.0242	+0.0007	[−0.0106, +0.0123]	1.03×	0.0250	0.0248	0.99×
4	0.0350	0.0328	−0.0023	[−0.0121, +0.0075]	0.94×	0.0368	0.0351	0.95×
8	0.0359	0.0345	−0.0014	[−0.0090, +0.0061]	0.96×	0.0369	0.0361	0.98×
16	0.0327	0.0306	−0.0021	[−0.0060, +0.0013]	0.94×	0.0329	0.0321	0.97×
32	0.0284	0.0273	−0.0011	[−0.0026, +0.0002]	0.96×	0.0281	0.0285	1.01×

Table 31: Observed agreement does not bound the latent gap of the questions with an undefined within-run correlation. Under the additional assumptions stated below the pooled bound is small beside the measured gap on every design where it is defined; the per-question bound is small on Arms A and B but several times the measured gap on the headline design, so it is uninformative there. The ratio columns divide the per-question bound by the measured gap, and the pooled bound by the measured gap and by the lower end of its interval. Sensitivity of the unobserved gap on questions with undefined within-run correlation. Their measured gap is zero, but this does not imply a zero latent gap. Write $\eta_q = \mathbb{P}(\text{a run is not all-equal})$. This probability alone does not bound $g_q = \text{pass@}k_q^{\text{iid}} - \text{succ@}k_q$: an endpoint mixture can have $\eta_q = 0$ and $g_q > 0$. The displayed numbers require the additional assumptions $1 - \mu_{1q} \le \eta_q$ on the set $\mathcal{A}_1$ of observed all-success questions, and $\mu_{1q} \le \eta_q$ on the set $\mathcal{A}_0$ of observed all-failure questions. Under the latent-mixture model, $g_q \le 1 - \mu_{1q}$ and $g_q \le (k-1)\mu_{1q}$; the extra assumptions, not observed agreement, link these inequalities to η_q. Requiring each selected question's observed-agreement event to have probability at least $\alpha = 0.05$ yields $\eta_q \le 1 - \alpha^{1/R}$. The pooled column additionally imposes a common η and independence across the selected questions, and requires their joint agreement event to have probability at least α, yielding $\eta \le 1 - \alpha^{1/(R|\mathcal{A}|)}$, with $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_0$. Both columns are therefore sensitivity values for a data-selected set, not unconditional bounds or confidence limits. Their target is the selected questions' contribution to the full-corpus $\text{pass@}k^{\text{iid}} - \text{succ@}k$ gap: $(|\mathcal{A}_1| + (k-1)|\mathcal{A}_0|)\eta/Q$; the measured comparator is $\widehat{U}_k - \overline{Y}$. The computation's branch-by-branch checks use the probabilities of not all-success and not all-failure instead; they do not validate a bound based on η_q alone.

k	per-question bound	pooled bound	measured $\widehat{U}_k - \overline{Y}$	interval lower end	ratio to gap (sep.)	ratio to gap (pool.)	ratio to lower end
Headline design: GSM8K, 7B model: $R = 16$, $n = 8$, 200 questions							
135 *all-success,* 3 *all-failure;* $\eta \le 0.1707$ *per question,* $\eta \le 0.001356$ *pooled*							
2	0.1178	0.000936	+0.0132	+0.0067	8.958×	0.071×	0.139×
4	0.1229	0.000976	+0.0262	+0.0156	4.701×	0.037×	0.063×
8	0.1332	0.001058	+0.0266	+0.0148	4.999×	0.040×	0.072×
16	0.1537	0.001220	n/a	n/a	n/a	n/a	n/a
32	0.1947	0.001546	n/a	n/a	n/a	n/a	n/a
$\hat{\rho}$ *quartiles where defined (*62 *questions):* 0.016 / 0.221 / 0.649							
*the same with the undefined questions entered at zero (*200*):* 0.000 / 0.000 / 0.000							
Arm A: GSM8K, 1.5B model: $R = 32$, $n = 32$, 194 questions							
2 *all-success,* 0 *all-failure;* $\eta \le 0.0894$ *per question,* $\eta \le 0.045730$ *pooled*							
2	0.0009	0.000471	+0.0976	+0.0866	0.009×	0.005×	0.005×
4	0.0009	0.000471	+0.1678	+0.1480	0.005×	0.003×	0.003×
8	0.0009	0.000471	+0.2107	+0.1872	0.004×	0.002×	0.003×
16	0.0009	0.000471	+0.2322	+0.2048	0.004×	0.002×	0.002×
32	0.0009	0.000471	+0.2410	+0.2121	0.004×	0.002×	0.002×
$\hat{\rho}$ *quartiles where defined (*192 *questions):* 0.386 / 0.663 / 0.880							
*the same with the undefined questions entered at zero (*194*):* 0.380 / 0.649 / 0.877							
Arm B: MATH-500, 7B model: $R = 32$, $n = 32$, 64 questions							
0 *all-success,* 0 *all-failure;* $\eta \le 0.0894$ *per question, no such question pooled*							
2	0.0000	n/a	+0.0256	+0.0112	0.000×	n/a	n/a
4	0.0000	n/a	+0.0360	+0.0174	0.000×	n/a	n/a
8	0.0000	n/a	+0.0366	+0.0184	0.000×	n/a	n/a
16	0.0000	n/a	+0.0324	+0.0150	0.000×	n/a	n/a
32	0.0000	n/a	+0.0287	+0.0135	0.000×	n/a	n/a
$\hat{\rho}$ *quartiles where defined (*64 *questions):* 0.021 / 0.075 / 0.211							
*the same with the undefined questions entered at zero (*64*):* 0.021 / 0.075 / 0.211							

Table 32: On laws fitted to the headline design, run-level intervals cover the exact succ@k in nearly every trial, while the attempt-level bootstrap covers it less often than it contains the noisy held-out reference. What the containment measurement of Table 10 would have said against an exactly known truth: its protocol — $R = 16$ runs of $n = 8$ attempts per question, budget $k = 4$, $\delta = 0.05$, an eight-run estimation half against an eight-run held-out half, 50 splits per question — run on laws we choose, so the truth is available in closed form. Each question's law matches that question's first two moments: a Beta(a, b) where the fitted variance is interior (44 questions), a two-point law on $\{0, 1\}$ where it attains its upper bound $\mu(1 - \mu)$ (4), and a point mass where it is zero or undefined (152), for 200 questions in all. Columns: containment of the held-out reference $\overline{Y}$, the average over the held-out runs, which is itself an eight-run estimate and therefore noisy; coverage of the exact succ@k and of the exact pass@k^{iid} of the law; the correction, coverage of the exact succ@k minus containment of the held-out $\overline{Y}$ (the second column minus the first); mean width; and zero width, the fraction of intervals whose two endpoints coincide. The first two blocks fit one law per question to the headline design, so their held-out column is directly comparable to the measured 0.864 and 0.966 of Table 10; they differ only in whether the 50 splits reuse one realised design or draw a fresh one, that is, in whether they carry the sampling variation of the real protocol. The last two blocks give every question the same law, to show how far the correction moves when the questions stop being heterogeneous.

Interval	fraction of intervals containing			correction	mean width	zero width
	held-out $\overline{Y}$	exact succ@k	exact pass@k^{iid}			
one law fitted per question; the 50 *splits reuse one realised design*						
attempt-level bootstrap	0.862	0.768	0.846	−0.094	0.027	0.773
run-level RUNCS	0.975	1.000	0.958	+0.025	0.476	0.000
run-level Clopper–Pearson	0.970	0.998	0.945	+0.028	0.401	0.000
one law fitted per question; each split draws a fresh design						
attempt-level bootstrap	0.868	0.775	0.855	−0.093	0.029	0.780
run-level RUNCS	0.970	1.000	0.960	+0.030	0.476	0.000
run-level Clopper–Pearson	0.966	0.997	0.946	+0.031	0.401	0.000
one law for every question, matched to the shared-prefix regime						
attempt-level bootstrap	0.642	0.035	0.650	−0.607	0.002	0.109
run-level RUNCS	0.908	1.000	0.895	+0.093	0.459	0.000
run-level Clopper–Pearson	0.908	0.992	0.895	+0.085	0.386	0.000
one law for every question, Beta$(2, 2)$						
attempt-level bootstrap	0.260	0.506	0.771	+0.246	0.134	0.000
run-level RUNCS	0.781	0.998	0.986	+0.218	0.606	0.000
run-level Clopper–Pearson	0.760	0.986	0.915	+0.227	0.510	0.000

Table 33: On laws fitted to Arms A–C, run-level RUNCS covers the exact succ@k in nearly every trial, while the attempt-level bootstrap covers it in fewer than a third of trials in every block. What the containment measurement of Table 20 would have said against an exactly known truth, one calibration per arm: its protocol — $R = 32$ runs of $n = 32$ attempts per question, budgets $k = 4$ and $k = 16$, $\delta = 0.05$, a sixteen-run estimation half against a sixteen-run held-out half, 50 splits per question; 194 questions on Arm A, 64 on Arm B and 200 on Arm C — run on laws we choose, so the truth is available in closed form. Each question's law matches its first two moments in that arm: a Beta(a, b) where the fitted variance is interior (180, 56 and 172 questions on Arms A, B and C), a two-point law on $\{0, 1\}$ where it attains its upper bound $\mu(1 - \mu)$ (3, 1 and 9), and a point mass where it is zero or undefined (11, 7 and 19). The arms share every design parameter and differ only in these laws; the correction is a function of the law and not of the protocol, so each arm needs its own calibration. Arm C keeps every prefix its generation draws, so its fitted laws include the degenerate ones the admissibility condition removes on Arms A and B; its attempt-level correction is -0.059 at $k = 4$ and -0.099 at $k = 16$, of the same sign as Arm A's at both budgets. The columns are those of Table 32: containment of the held-out $\overline{Y}$, a sixteen-run estimate; coverage of the exact succ@k and pass@k^{iid} of the law; the correction, the second column minus the first; mean width; and zero width. The two blocks at each budget differ only in whether the 50 splits reuse one realised design or draw a fresh one. The numbers of Table 32, for the smaller design of Table 10, do not transfer here: the correction depends on the design.

Interval	fraction of intervals containing			correction	mean width	zero width
	held-out $\overline{Y}$	exact succ@k	exact pass@k^{iid}			
Arm A: GSM8K, 1.5B model						
budget $k = 4$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.195	0.135	0.363	−0.060	0.051	0.099
run-level RUNCS	0.900	0.999	0.612	+0.099	0.488	0.000
run-level Clopper–Pearson	0.856	0.990	0.488	+0.133	0.375	0.000
budget $k = 4$; each split draws a fresh design						
attempt-level bootstrap	0.192	0.132	0.355	−0.060	0.052	0.096
run-level RUNCS	0.904	0.998	0.601	+0.095	0.492	0.000
run-level Clopper–Pearson	0.855	0.984	0.490	+0.129	0.378	0.000
budget $k = 16$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.232	0.118	0.492	−0.114	0.045	0.204
run-level RUNCS	0.896	0.999	0.372	+0.102	0.479	0.000
run-level Clopper–Pearson	0.858	0.980	0.333	+0.122	0.368	0.000
budget $k = 16$; each split draws a fresh design						
attempt-level bootstrap	0.226	0.116	0.473	−0.110	0.045	0.213
run-level RUNCS	0.893	0.998	0.346	+0.105	0.483	0.000
run-level Clopper–Pearson	0.849	0.981	0.303	+0.132	0.371	0.000
Arm B: MATH-500, 7B model						
budget $k = 4$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.222	0.269	0.624	+0.047	0.029	0.041
run-level RUNCS	0.924	1.000	0.925	+0.076	0.376	0.000
run-level Clopper–Pearson	0.903	0.995	0.870	+0.092	0.289	0.000
budget $k = 4$; each split draws a fresh design						
attempt-level bootstrap	0.217	0.290	0.637	+0.073	0.030	0.053
run-level RUNCS	0.921	0.999	0.930	+0.078	0.377	0.000
run-level Clopper–Pearson	0.896	0.984	0.875	+0.088	0.290	0.000
budget $k = 16$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.466	0.187	0.719	−0.279	0.015	0.279
run-level RUNCS	0.962	1.000	0.883	+0.038	0.326	0.000
run-level Clopper–Pearson	0.948	0.996	0.869	+0.047	0.251	0.000
budget $k = 16$; each split draws a fresh design						
attempt-level bootstrap	0.451	0.199	0.739	−0.252	0.017	0.281
run-level RUNCS	0.946	0.999	0.874	+0.053	0.326	0.000
run-level Clopper–Pearson	0.934	0.995	0.853	+0.061	0.251	0.000

continued on the next page

Table 33, continued.

Interval	fraction of intervals containing			correction	mean width	zero width
	held-out $\overline{Y}$	exact succ@k	exact pass@k^{iid}			
Arm C: GSM8K, 1.5B model, no prefix admissibility condition						
budget $k = 4$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.224	0.165	0.372	−0.059	0.053	0.150
run-level runCS	0.920	0.998	0.611	+0.077	0.496	0.000
run-level Clopper–Pearson	0.862	0.976	0.478	+0.114	0.381	0.000
budget $k = 4$; each split draws a fresh design						
attempt-level bootstrap	0.219	0.159	0.369	−0.061	0.053	0.133
run-level runCS	0.917	0.998	0.585	+0.081	0.498	0.000
run-level Clopper–Pearson	0.871	0.983	0.468	+0.111	0.382	0.000
budget $k = 16$; the 50 splits reuse one realised design						
attempt-level bootstrap	0.242	0.144	0.486	−0.099	0.043	0.237
run-level runCS	0.910	0.997	0.326	+0.086	0.489	0.000
run-level Clopper–Pearson	0.869	0.980	0.298	+0.112	0.376	0.000
budget $k = 16$; each split draws a fresh design						
attempt-level bootstrap	0.229	0.135	0.488	−0.093	0.042	0.234
run-level runCS	0.915	0.998	0.329	+0.083	0.492	0.000
run-level Clopper–Pearson	0.866	0.979	0.302	+0.113	0.378	0.000

Table 34: At budget 4, in every design, the question-level interval for the attempt-level bootstrap's containment lies below the interval for every run-level construction's containment. The intervals are percentile bootstraps that resample questions and carry all of a question's splits with them, the question being the independent unit (the splits of a question reuse the same runs, so a binomial standard error over intervals would be too small). Columns: containment of the held-out reference, and zero width, the fraction of intervals whose endpoints coincide, each with its interval; the point figures are those of Tables 10, 20 and 25. Every budget: Table S14 of the supplementary report.

Interval	Containment	95% CI	Zero width	95% CI
Headline design: GSM8K, 7B model, budget $k = 4$, 200 questions, 10,000 intervals				
attempt-level bootstrap	0.864	[0.822, 0.906]	76.9%	[71.3%, 82.4%]
run-level runCS	0.966	[0.950, 0.980]	0%	[0%, 0%]
run-level Clopper–Pearson	0.960	[0.944, 0.974]	0%	[0%, 0%]
run-level bootstrap	0.957	[0.941, 0.973]	86.1%	[81.5%, 90.3%]
run-level bootstrap, factor removed	0.957	[0.940, 0.973]	86.1%	[81.5%, 90.4%]
run-level Wald, unbiased variance	0.947	[0.929, 0.965]	86.1%	[81.8%, 90.4%]
Arm A: GSM8K, 1.5B model, budget $k = 4$, 194 questions, 9,700 intervals				
attempt-level bootstrap	0.146	[0.111, 0.185]	6.9%	[4.4%, 9.8%]
run-level runCS	0.892	[0.871, 0.911]	0%	[0%, 0%]
run-level Clopper–Pearson	0.838	[0.818, 0.856]	0%	[0%, 0%]
run-level bootstrap	0.841	[0.821, 0.860]	17.1%	[12.9%, 21.6%]
run-level bootstrap, factor removed	0.841	[0.822, 0.861]	17.1%	[12.8%, 21.8%]
run-level Wald, unbiased variance	0.781	[0.761, 0.802]	17.1%	[12.9%, 21.5%]
Arm B: MATH-500, 7B model, budget $k = 4$, 64 questions, 3,200 intervals				
attempt-level bootstrap	0.208	[0.136, 0.288]	4.8%	[0.8%, 10.2%]
run-level runCS	0.922	[0.883, 0.958]	0%	[0%, 0%]
run-level Clopper–Pearson	0.902	[0.862, 0.938]	0%	[0%, 0%]
run-level bootstrap	0.900	[0.864, 0.933]	59.4%	[48.0%, 70.9%]
run-level bootstrap, factor removed	0.900	[0.862, 0.936]	59.4%	[48.1%, 70.1%]
run-level Wald, unbiased variance	0.875	[0.834, 0.913]	59.4%	[48.1%, 70.1%]

continued on the next page

Table 34, continued.

Interval	Containment	95% CI	Zero width	95% CI
Arm C: GSM8K, 1.5B model, no prefix admissibility condition, budget $k = 4$, 200 questions, 10,000 intervals				
attempt-level bootstrap	0.190	[0.148, 0.235]	11.3%	[7.6%, 15.1%]
run-level runCS	0.905	[0.885, 0.923]	0%	[0%, 0%]
run-level Clopper–Pearson	0.858	[0.839, 0.875]	0%	[0%, 0%]
run-level bootstrap	0.857	[0.840, 0.874]	19.5%	[14.8%, 24.5%]
run-level bootstrap, factor removed	0.857	[0.839, 0.875]	19.5%	[15.2%, 24.3%]
run-level Wald, unbiased variance	0.807	[0.789, 0.825]	19.5%	[15.0%, 24.3%]
Self-debugging loop, 7B model, budget $k = 4$, 164 questions, 8,200 intervals				
attempt-level bootstrap	0.785	[0.736, 0.832]	55.8%	[48.5%, 63.0%]
run-level runCS	0.925	[0.900, 0.948]	0%	[0%, 0%]
run-level Clopper–Pearson	0.920	[0.896, 0.945]	0%	[0%, 0%]
run-level bootstrap	0.919	[0.893, 0.943]	82.2%	[76.9%, 86.8%]
run-level bootstrap, factor removed	0.919	[0.892, 0.944]	82.2%	[77.1%, 86.9%]
run-level Wald, unbiased variance	0.912	[0.885, 0.938]	82.2%	[76.8%, 87.3%]
Self-debugging loop, 1.5B model, budget $k = 4$, 164 questions, 8,200 intervals				
attempt-level bootstrap	0.670	[0.624, 0.713]	27.4%	[21.6%, 33.3%]
run-level runCS	0.887	[0.862, 0.908]	0%	[0%, 0%]
run-level Clopper–Pearson	0.873	[0.850, 0.898]	0%	[0%, 0%]
run-level bootstrap	0.860	[0.836, 0.886]	47.2%	[40.5%, 54.2%]
run-level bootstrap, factor removed	0.860	[0.834, 0.885]	47.2%	[40.5%, 53.9%]
run-level Wald, unbiased variance	0.814	[0.786, 0.841]	47.2%	[40.6%, 53.8%]

Table 35: On laws fitted to the reasoning tree, run-level intervals cover the exact succ@k in nearly every trial, so reference noise explains their measured shortfall on those laws. What the containment measurement on the reasoning tree would have said against an exactly known truth. The protocol is that of the tree's containment rows in Table 16: every question keeps its own number R_q of sibling groups, unbalanced, as in the corpus — 4,880 questions with R_q between 4 and 23, median 7, only questions with $R_q \geq 4$ entering — each group carrying $n = 4$ siblings, an $R_q/2$ estimation half against an $R_q/2$ held-out half, budgets $k = 2$ and $k = 4$, $\delta = 0.05$, 20 splits per question. It is run on laws chosen to match each question's two moments, so the truth is available in closed form: a Beta(a, b) where the fitted variance is interior (4,175 questions), a two-point law on $\{0, 1\}$ where it attains its upper bound $\mu(1-\mu)$ (35 questions), and a point mass where it is zero or undefined (670 questions). Columns: the containment measured on the real corpus; the same containment reproduced on the fitted laws, of the held-out $\overline{Y}$, an $R_q/2$ estimate and therefore noisy; coverage of the exact succ@k and pass@k^{iid} of the law; the correction, the third column minus the second; mean width; and zero width, the fraction of intervals whose two endpoints coincide. The largest gap between the first two columns is 0.015, which is the accuracy of a two-moment fit and the reason the correction is read as a shift and not as a replacement. The blocks at each budget differ only in whether the 20 splits reuse one realised design or draw a fresh one. Tables 32 and 33 do the same for the shared-prefix designs, and their numbers do not transfer here: the correction depends on the design.

Interval	measured	fraction of intervals containing		correction	mean width	zero width
		held-out $\overline{Y}$	exact succ@k	exact pass@k^{iid}			
budget $k = 2$; the 20 splits reuse one realised design							
attempt-level bootstrap	0.629	0.643	0.734	0.754	+0.091	0.478	0.109
run-level RUNCS	0.785	0.785	0.998	0.995	+0.213	0.854	0.000
run-level Clopper–Pearson	0.768	0.770	0.992	0.986	+0.222	0.799	0.000
budget $k = 2$; each split draws a fresh design							
attempt-level bootstrap	0.629	0.644	0.737	0.756	+0.093	0.479	0.110
run-level RUNCS	0.785	0.786	0.998	0.995	+0.212	0.855	0.000
run-level Clopper–Pearson	0.768	0.772	0.993	0.987	+0.221	0.799	0.000
budget $k = 4$; the 20 splits reuse one realised design							
attempt-level bootstrap	0.609	0.608	0.640	0.740	+0.032	0.457	0.133
run-level RUNCS	0.788	0.795	0.998	0.967	+0.203	0.845	0.000
run-level Clopper–Pearson	0.772	0.780	0.994	0.932	+0.213	0.790	0.000
budget $k = 4$; each split draws a fresh design							
attempt-level bootstrap	0.609	0.613	0.643	0.741	+0.029	0.461	0.133
run-level RUNCS	0.788	0.795	0.998	0.969	+0.203	0.845	0.000
run-level Clopper–Pearson	0.772	0.780	0.993	0.931	+0.213	0.790	0.000

Table 36: Effective attempts and the threshold $R \geq k$ order designs in opposite directions (budget $k = 4$). The effective attempt count $n/[1 + (n-1)\bar{\rho}]$, with $\bar{\rho}$ the average within-run correlation, scores a design by the precision its attempts can reach; the threshold $R \geq k$ scores it by whether the quantity is estimable without bias at all. The first six rows hold the total attempt count fixed at 32; the last two are the pair that separates the two readings, one with more total effective attempts and no unbiased estimator, the other with fewer and one. The bias column is the finite-design expectation of the pooled estimator, Proposition A.21, evaluated at that design with the moments of the full sample, minus the full-sample $\widehat{U}_k$: an estimate of $\mathbb{E}[\widehat{C}_{\text{pooled}}] - \text{pass@}k^{\text{iid}}$. Negative values mean the pooled report falls below the independent-attempt target $\text{pass@}k^{\text{iid}}$; its excess over succ@$k$ is a different comparison (Table 1). The standard error column is the exact variance of $\widehat{U}_k$ under that design, aggregated over questions. At $n = 1$ the pooled estimator is $\widehat{U}_k$ itself and its expectation is exactly $\text{pass@}k^{\text{iid}}$ (Corollary A.23), so the entries of the $n = 1$ rows are the error of evaluating that expectation with estimated moments, and they set the resolution of the bias column. No design with n larger than the generating design is shown: the moments $m_c = \mathbb{E}_\pi[(1-\theta)^c]$ for c beyond n are not estimable from a run of n attempts, and Theorem 4.1(b) says that gap does not close by sampling harder.

$(R,\ n)$	attempts	effective attempts		$R \geq k$	$\mathbb{E}[\widehat{C}_{\text{pooled}}] - \widehat{U}_k$		SE of $\widehat{U}_k$
		per run	total		mean	95% CI	
Arm A: GSM8K, 1.5B model: $\bar{\rho} = 0.6838$, full design $R = 32$, $n = 32$, 194 questions							
(1, 32)	32	1.442	1.442	**no**	−0.1690	[−0.1881, −0.1511]	n/a
(2, 16)	32	1.421	2.843	**no**	−0.0790	[−0.0893, −0.0685]	n/a
(4, 8)	32	1.382	5.530	yes	−0.0377	[−0.0434, −0.0324]	0.01401
(8, 4)	32	1.311	10.487	yes	−0.0184	[−0.0211, −0.0155]	0.00949
(16, 2)	32	1.188	19.004	yes	−0.0091	[−0.0104, −0.0078]	0.00716
(32, 1)	32	1.000	32.000	yes	−0.0046	[−0.0053, −0.0038]	0.00583
(3, 32)	96	1.442	4.325	**no**	−0.0533	[−0.0604, −0.0460]	n/a
(4, 1)	4	1.000	4.000	yes	−0.0046	[−0.0053, −0.0038]	0.01984
Arm B: MATH-500, 7B model: $\bar{\rho} = 0.2007$, full design $R = 32$, $n = 32$, 64 questions							
(1, 32)	32	4.431	4.431	**no**	−0.0388	[−0.0572, −0.0233]	n/a
(2, 16)	32	3.990	7.980	**no**	−0.0173	[−0.0257, −0.0100]	n/a
(4, 8)	32	3.327	13.307	yes	−0.0081	[−0.0124, −0.0048]	0.01038
(8, 4)	32	2.497	19.974	yes	−0.0040	[−0.0058, −0.0022]	0.00845
(16, 2)	32	1.666	26.651	yes	−0.0020	[−0.0030, −0.0012]	0.00760
(32, 1)	32	1.000	32.000	yes	−0.0011	[−0.0016, −0.0006]	0.00718
(3, 32)	96	4.431	13.294	**no**	−0.0116	[−0.0166, −0.0070]	n/a
(4, 1)	4	1.000	4.000	yes	−0.0011	[−0.0015, −0.0006]	0.02576

Table 37: Starting a run costs a few attempts' worth of generation on this paper's own runs, with a wide interval, and the allocation it implies depends on the mixing law. The cost ratio $\kappa = c_0/c_1$ of Proposition A.24 estimated from this paper's own generation runs, and the allocation of attempts it implies. Panel (a): each question's wall clock is regressed on the tokens generated in its two segments, $t_q \approx a p_q + b c_q + d$, where p_q counts every prefix candidate drawn including the rejected ones and c_q counts the continuation tokens; $c_0 = aP \cdot$ (candidates per accepted prefix) is then the cost of one admissible prefix and $c_1 = b\,\bar{\ell}$ the cost of one continuation, with $\bar{\ell}$ the mean number of tokens per continuation. Intervals are percentile bootstrap over questions, 2000 resamples, the question being the independent unit. The slope a is identified only where the number of oversampling rounds varies across questions, which happens on the GSM8K 1.5B row (Arm A) only; on the MATH-500 rows every question finished in one round, so a is collinear with the intercept and only the token proxy $\kappa_{\text{tok}} = P \cdot$ (candidates per accepted prefix)$/\bar{\ell}$ is available. The last column reads the same runs as average rather than marginal cost, assigning the per-question fixed overhead entirely to the continuation segment, and is therefore the lower end of the range. An independent probe agrees with the fitted a to within 0.083 in relative terms: the 6 questions that exhausted the oversampling budget without filling their prefix quota generated prefixes and no continuations at all, so their wall clock divided by their prefix tokens estimates a without sharing any information with the regression. Panel (b): $n^\star$ recomputed at those estimates on the four mixing laws of Table 5, through the same closed form; the first four columns reproduce that table's assumed grid cell for cell, which is what licenses reading the remaining columns against it. The MATH-500 row "earlier pass" times the questions that an earlier, partial generation pass produced and the main MATH-500 pass reused, so their time is counted there and not in the main row. The questions column counts the questions whose wall clock enters the fit. On Arm A 188 of the 194 analysed questions enter: the other 6 are each timed together with one of the 6 excluded questions just before them. The main MATH-500 pass timed 48 questions and leaves out one, whose time includes copying the 16 reused ones; with the 16 of the earlier pass, 63 of Arm B's 64 questions enter.

(a) estimated cost of a prefix and of a continuation, in seconds

Arm	questions	a (s/token)	b (s/token)	c_0 (s)	c_1 (s)	$\hat{\kappa}$	95% CI	κ_{tok}	κ amortised
GSM8K, 1.5B	188	3.92×10^{-4}	2.90×10^{-4}	0.0466	0.0158	2.96	[0.18, 6.32]	2.19	1.15
MATH-500, 7B	47	n/a	1.04×10^{-3}	n/a	0.3988	n/a	n/a	0.25	n/a
MATH-500, 7B (earlier pass)	16	n/a	8.77×10^{-4}	n/a	0.2783	n/a	n/a	0.30	n/a

(b) $n^\star$ of Proposition A.24 at budget $k = 4$

	at the assumed grid of Table 5				at the estimate		
Mixing law (μ_1, Var(θ))	$\kappa = 0.1$	$\kappa = 1$	$\kappa = 10$	$\kappa = 100$	CI low	$\hat{\kappa}$	CI high
two-point (0.90, 0.020)	28	30	40	88	28	32	37
two-point (0.38, 0.068)	4	4	4	7	4	4	4
Beta (0.90, 0.020)	7	8	12	30	7	9	11
Beta (0.38, 0.068)	4	4	8	20	4	5	7

C.3 The decision experiment

This subsection holds the decision experiment of §F.6, two systems compared with eight interval procedures: Table 38 gives, for every simulated design with at most 16 runs, the cells the main text quotes, and Table 39 the same experiment on measured runs. The full grids, with every procedure and every design, are in Supplement §S4.

Table 38: At a fixed question set one run's attempt-level bootstrap certifies a difference between identical systems in most replicates, while the cluster-robust Wald interval keeps its type I near the nominal 0.05, with power 0.955–0.968 at $R = 8$, and run-level runCS has power at most 0.001 at eight runs or fewer. One row per simulated design with at most 16 runs, the rows every decision-error range of the main text is taken over. Two systems are compared whose question-averaged succ@4 differs by 0.050 (the alternative) or not at all (the null); frame (a) resamples questions and frame (b) holds the question set fixed; κ is the cost of starting a run in attempts, N the total cost, R runs of n attempts. Columns: the type I rate of the attempt-level bootstrap and of the cluster-robust Wald interval; the power (certifying the true difference with its sign) of the Wald interval, of run-level runCS and of the discordant-pair runCS, which certifies the sign only. "n/a": undefined by construction (fewer than k attempts per run, or one run). The full grids — all eight procedures, wrong-sign rates, widths and the designs with more runs — are Tables S22, S23 and S24 of the supplementary report.

					type I		power		
frame	κ	N	R	n	attempt	Wald	Wald	runCS	disc. runCS
equal attempts, $\kappa = 0$ (full grid: Table S22)									
(a)	0	32	1	32	0.050	n/a	n/a	0.000	0.029
(a)	0	32	2	16	0.043	0.060	0.470	0.000	0.067
(a)	0	32	4	8	0.038	0.060	0.715	0.000	0.179
(a)	0	32	8	4	0.021	0.049	0.956	0.000	0.551
(a)	0	128	1	128	0.044	n/a	n/a	0.000	0.028
(a)	0	128	2	64	0.052	0.044	0.456	0.000	0.068
(a)	0	128	4	32	0.050	0.051	0.731	0.000	0.198
(a)	0	128	8	16	0.046	0.063	0.961	0.001	0.573
(b)	0	32	1	32	0.745	n/a	n/a	0.000	0.029
(b)	0	32	2	16	0.618	0.060	0.479	0.000	0.067
(b)	0	32	4	8	0.428	0.057	0.727	0.000	0.179
(b)	0	32	8	4	0.238	0.047	0.964	0.000	0.551
(b)	0	128	1	128	0.865	n/a	n/a	0.000	0.028
(b)	0	128	2	64	0.802	0.045	0.467	0.000	0.068
(b)	0	128	4	32	0.668	0.052	0.738	0.000	0.198
(b)	0	128	8	16	0.531	0.066	0.966	0.001	0.573
equal attempts, $\kappa = 0$, further designs (Table S23)									
(a)	0	32	16	2	0.012	n/a	n/a	n/a	n/a
(a)	0	128	16	8	0.033	0.051	1.000	0.026	0.959
(b)	0	32	16	2	0.118	n/a	n/a	n/a	n/a
(b)	0	128	16	8	0.356	0.047	1.000	0.026	0.959
equal cost, $\kappa = 2.96$ (Table S24, panels (a)–(b))									
(a)	2.96	32	1	29	0.045	n/a	n/a	0.000	0.023
(a)	2.96	32	2	13	0.048	0.054	0.447	0.000	0.070
(a)	2.96	32	4	5	0.028	0.051	0.730	0.000	0.203
(a)	2.96	32	8	1	0.007	n/a	n/a	n/a	n/a
(a)	2.96	128	1	125	0.051	n/a	n/a	0.000	0.023
(a)	2.96	128	2	61	0.049	0.064	0.478	0.000	0.073
(a)	2.96	128	4	29	0.048	0.043	0.731	0.000	0.189
(a)	2.96	128	8	13	0.039	0.042	0.955	0.001	0.577
(a)	2.96	128	16	5	0.029	0.056	0.999	0.030	0.944
(b)	2.96	32	1	29	0.741	n/a	n/a	0.000	0.023
(b)	2.96	32	2	13	0.579	0.053	0.461	0.000	0.070
(b)	2.96	32	4	5	0.327	0.049	0.737	0.000	0.203
(b)	2.96	32	8	1	0.069	n/a	n/a	n/a	n/a

continued on the next page

Table 38, continued.

frame	κ	N	R	n	attempt	Wald	Wald	runCS	disc. runCS
(b)	2.96	128	1	125	0.855	n/a	n/a	0.000	0.023
(b)	2.96	128	2	61	0.789	0.059	0.482	0.000	0.073
(b)	2.96	128	4	29	0.645	0.041	0.749	0.000	0.189
(b)	2.96	128	8	13	0.465	0.041	0.961	0.001	0.577
(b)	2.96	128	16	5	0.267	0.053	1.000	0.030	0.944
total cost 512 *(Table S24, panels (c)–(d))*									
(a)	0	512	1	512	0.064	n/a	n/a	0.000	0.028
(a)	0	512	2	256	0.063	0.047	0.449	0.000	0.067
(a)	0	512	4	128	0.046	0.055	0.730	0.000	0.207
(a)	0	512	8	64	0.044	0.046	0.961	0.001	0.577
(a)	0	512	16	32	0.054	0.051	0.999	0.032	0.951
(a)	2.96	512	1	509	0.066	n/a	n/a	0.000	0.028
(a)	2.96	512	2	253	0.050	0.053	0.434	0.000	0.057
(a)	2.96	512	4	125	0.051	0.054	0.753	0.000	0.192
(a)	2.96	512	8	61	0.054	0.058	0.961	0.000	0.576
(a)	2.96	512	16	29	0.059	0.047	1.000	0.024	0.949
(b)	0	512	1	512	0.931	n/a	n/a	0.000	0.028
(b)	0	512	2	256	0.899	0.046	0.450	0.000	0.067
(b)	0	512	4	128	0.810	0.052	0.741	0.000	0.207
(b)	0	512	8	64	0.719	0.044	0.968	0.001	0.577
(b)	0	512	16	32	0.619	0.050	0.999	0.032	0.951
(b)	2.96	512	1	509	0.936	n/a	n/a	0.000	0.028
(b)	2.96	512	2	253	0.890	0.054	0.451	0.000	0.057
(b)	2.96	512	4	125	0.841	0.052	0.763	0.000	0.192
(b)	2.96	512	8	61	0.745	0.067	0.962	0.000	0.576
(b)	2.96	512	16	29	0.624	0.051	1.000	0.024	0.949

Table 39: On measured runs the attempt-level bootstrap's type I at a fixed question set tracks the simulated one design by design, and the cluster-robust Wald interval keeps its type I near the nominal level. The null splits the 32 runs of each Arm A question into two halves treated as two systems; the alternative is Arm A against Arm C on their 194 shared questions (the difference of the two held-out halves' succ@4 is the truth of a split, a smaller difference than the simulated one, so power is not comparable between the two tables). Frames, κ, N and the procedures are those of Table 38; R' runs of n' attempts are cut from each half. "sim." is the simulated attempt-level type I of the same design, the corresponding cell of Table 38; "n/a" marks an entry undefined by construction. The full grid, with all eight procedures, wrong-sign rates and widths, is Table S25 of the supplementary report.

					type I			power		
frame	κ	N	R'	n'	attempt	sim.	Wald	Wald	runCS	disc. runCS
(a)	0	32	1	32	0.035	0.050	n/a	n/a	0.000	0.003
(a)	0	32	2	16	0.063	0.043	0.040	0.105	0.000	0.008
(a)	0	32	4	8	0.038	0.038	0.060	0.138	0.000	0.005
(a)	0	32	8	4	0.038	0.021	0.060	0.278	0.000	0.010
(a)	0	32	16	2	0.008	0.012	n/a	n/a	n/a	n/a
(a)	0	128	4	32	0.058	0.050	0.060	0.138	0.000	0.005
(a)	0	128	8	16	0.043	0.046	0.060	0.278	0.000	0.010
(a)	0	128	16	8	0.025	0.033	0.035	0.503	0.000	0.020
(a)	2.96	32	1	29	0.038	0.045	n/a	n/a	0.000	0.003
(a)	2.96	32	2	13	0.063	0.048	0.040	0.105	0.000	0.008
(a)	2.96	32	4	5	0.028	0.028	0.060	0.138	0.000	0.005
(a)	2.96	32	8	1	0.015	0.007	n/a	n/a	n/a	n/a
(a)	2.96	128	4	29	0.053	0.048	0.060	0.138	0.000	0.005
(a)	2.96	128	8	13	0.040	0.039	0.060	0.278	0.000	0.010

continued on the next page

Table 39, continued.

frame	κ	N	R'	n'	attempt	sim.	Wald	Wald	runCS	disc. runCS
(a)	2.96	128	16	5	0.015	0.029	0.035	0.503	0.000	0.020
(b)	0	32	1	32	0.715	0.745	n/a	n/a	0.000	0.003
(b)	0	32	2	16	0.615	0.618	0.040	0.115	0.000	0.008
(b)	0	32	4	8	0.393	0.428	0.060	0.155	0.000	0.005
(b)	0	32	8	4	0.253	0.238	0.055	0.338	0.000	0.010
(b)	0	32	16	2	0.108	0.118	n/a	n/a	n/a	n/a
(b)	0	128	4	32	0.668	0.668	0.060	0.155	0.000	0.005
(b)	0	128	8	16	0.518	0.531	0.055	0.338	0.000	0.010
(b)	0	128	16	8	0.338	0.356	0.028	0.705	0.000	0.020
(b)	2.96	32	1	29	0.698	0.741	n/a	n/a	0.000	0.003
(b)	2.96	32	2	13	0.580	0.579	0.040	0.115	0.000	0.008
(b)	2.96	32	4	5	0.310	0.327	0.060	0.155	0.000	0.005
(b)	2.96	32	8	1	0.073	0.069	n/a	n/a	n/a	n/a
(b)	2.96	128	4	29	0.640	0.645	0.060	0.155	0.000	0.005
(b)	2.96	128	8	13	0.473	0.465	0.055	0.338	0.000	0.010
(b)	2.96	128	16	5	0.243	0.267	0.028	0.705	0.000	0.020

D Data and protocol details

This appendix records what we generated, what we obtained from public corpora, and where the artefacts and their manifests live in the released package.

D.1 The freshness checklist for Assumption 5.1

Assumption 5.1 is the one hypothesis of §5 that an evaluation harness can violate silently: nothing in the numbers announces that two runs shared sampled or mutable state; the run-level argument requires the resulting observations to remain independent. Table 40 is the operational form of the assumption — practical controls for isolating stochastic state between runs, distinguishing it from fixed read-only inputs and recording the reset policy.

D.2 How the number is computed in practice

Table 41 records, for ten recent evaluations of dependent procedures and for three public scoring harnesses, whether the attempts of a run share a sampled state, how the reported pass@k is computed, and whether it is set beside an any-success rate at $k \geq 2$ computed from independent samples; Table S28 reads the harnesses against Table 40. The audit is of what the sources state and what the code does, read in September 2026.

Three entries are worth naming here. Shared-decomposition sampling is explicit in Liang et al. (2026, Fig. 6); its pass@32 applies (3) to its candidates ("DAC" in that paper) without stating how many are scored per question or how they are allocated across decompositions, so pooling ($n > k$) is not established, and with $n = k$ (3) is the any-success indicator of the complete procedure. MapCoder and AlphaCodium (Islam et al., 2024; Ridnik et al., 2024) contain internal planning or revision, but this alone does not establish dependence between their outer pass@5 executions; historical outer-run freshness remains unverified.

Table 40: Freshness is checked by resetting or recording, per run, each source of stochastic state; a missing record leaves freshness unverified and does not itself prove dependence. Practical controls and records for Assumption 5.1. Fixed read-only inputs may be shared. The mathematical requirement is independent, identically distributed run observations.

State to examine	Potential shared-state effect	Practical control
Sampling seed	Reusing a seed with the same state can replicate outputs.	Seed recorded per run and asserted distinct.
Prompt cache and prefix / KV reuse	Reuse of sampled tokens or seeds across runs reintroduces the shared-prefix mechanism of §6.1 at the unit level; a cache of the deterministic prompt alone leaves the sampling law unchanged.	Sampled state not reused across runs; cache scope recorded in the run manifest.
Retrieval context and index state	Can correlate runs through documents retrieved on the basis of earlier runs.	Retrieval state rebuilt per run; index revision pinned.
Tool and environment state	Can correlate agent runs through files, containers or scratch state left by a previous run.	Fresh container or workspace per run, asserted at run start.

Work	dep. (Q1)	computed as (Q2)	beside indep., $k \geq 2$ (Q3)	where

Table 41: Six of the ten evaluations score attempts that share sampled state, and three of them set the number beside an any-success rate at $k \geq 2$ computed from independent samples. The audit covers how pass@k is computed and compared in dependent-procedure evaluations and in three public harnesses (public sources inspected September 2026; commit hashes for the harnesses in the Work column). "Attempts dependent at the scored unit?" answers Q1: do the executions counted by the score share sampled state? Internal steps and complete executions are distinguished. "How pass@k is computed" answers Q2 in words; the entries are of nine kinds: (a) $n \geq k$ independent draws scored with the Chen et al. (2021) estimator (3); (b) dependent attempts pooled into one bag and the same estimator applied to the pool; (c) a per-run indicator of success within the first k attempts, averaged over runs or questions – the succ@k-type quantity of §5, whether or not the source labels it that way; (d) only pass@1 is reported; (e) no pass@k-style label is used at all; (f) the score counts complete executions, whose independence is not established by the cited inner flow; (g) the same estimator applied to dependent candidates without the number scored per question being stated, so whether dependent attempts are pooled ($n > k$) is not established; (h) the same estimator on $n \geq k$ independent complete runs; (i) for a harness, the same estimator applied to whatever samples the caller supplies, with their independence not checked. An entry of unknown is not evidence of either dependence or independence. "Set beside an independent any-success rate at $k \geq 2$?" answers Q3: is the number placed in the same table or figure, in the same column or on the same axis, as an any-success rate at $k \geq 2$ (pass@k or oracle best-of-k) computed from independent samples? A pass@1 comparison does not count: at $k = 1$ the procedure's success probability and the independent-attempt one coincide. For the three harnesses Q1 and Q3 are marked n/a: a harness scores whatever samples its caller supplies, it does not itself generate a dependent procedure.

Work	dep. (Q1)	computed as (Q2)	beside indep., $k \geq 2$ (Q3)	where
literature: dependent-procedure evaluations				
(Islam et al., 2024)	unknown	complete executions; outer independence unverified	unknown	Original Algorithm1 inner k is not evaluation pass@k; official Base.py complete run_single_pass loop and MapCoder.py fresh exemplars/plans/code per invocation
(Liang et al., 2026)	yes	Chen estimator on dependent candidates; number scored not stated, pooling unverified	yes	§2.1 (division/conquering def.); §3.1 evaluation (Pass@1 averaged over 32 runs; Pass@32 with the Chen et al. 2021 estimator; candidate construction not stated); Table 1; Fig. 6 (shared-decomposition allocations beside 1024 independent CoT generations: Q1, Q3)
(Shinn et al., 2023)	yes	pass@1 only	no	Algorithm 1 p.4 (mem accumulates reflections across trials); Table 1 p.7 ("Pass@1" column); Table 4 p.12 (averaged over eight trials)
(Yao et al., 2023)	yes	no pass@k label	yes	§3 p.3 (tree branches, shared prefix); Table 2 p.6; p.6 oracle best-of-k; Table 7 / Appendix p.14

Work	dep. (Q1)	computed as (Q2)	beside indep., $k \geq 2$ (Q3)	where
(Chen et al., 2024)	yes	pass@1 only	yes	§2/Fig. 1; §4.2 (all TransCoder unit tests in the prompt); §5 (greedy initial program, max 10 turns); Table 2(b),(c) ("Oracle Baseline": one further independent sample when greedy fails, same column as the self-debugging rows); App. D Table 6 ("TransCoder (pass@1)"); Fig. 5a vs. 5b, App. E.1 (ICLR 2024 version)
(Zhou et al., 2024)	yes	pass@1 only	no	§4 (MCTS, shared value fn/ reflections); §5.2 p.6-7 (k=8 iterations × n=5 solutions/ expansion, select highest-value); Table 4/5 ("Pass@1")
(Ridnik et al., 2024)	un known	complete executions; outer independence unverified	un known	Section 5.3 counts calls per final solution; official solve_problem_ in_dataset rebuilds each problem and flow
(Brown et al., 2024)	no	$n \geq k$ independent draws, Chen estimator	no	§2 p.2,4 (Eq.1, Chen et al. cite verbatim); p.5/p.12 SWE-bench Appendix ("all attempts made independently of one another")
(Bjarnason et al., 2026)	no	$n \geq k$ independent complete runs, Chen estimator	no	§2.3 p.4 Eq. (3) (pass@k / pass^k over m independent runs); §3.2 p.9-10 (Recommendations)
(Snell et al., 2025)	yes	pass@1 only	no	§3 p.4 (parallel independent vs. sequential revision contrast); Fig. 6 (left: pass@1 per revision step; right: selected-answer accuracy)

Work	dep. (Q1)	computed as (Q2)	beside indep., $k \geq 2$ (Q3)	where
public harnesses (Q1/Q3 not applicable: a harness scores whatever samples it is given, it does not itself generate a dependent procedure)				
bigcode-evaluation-harness (`8fc5bae6`)	n/a	Chen estimator applied to caller-supplied samples; independence not checked	n/a	bigcode_eval/tasks/custom_metrics/code_eval.py L129-188 (compute_code_eval / estimate_pass_at_k, verbatim OpenAI HumanEval formula); arguments.py L36-37; generation.py L65; utils.py L277-305; tasks/custom_metrics/execute.py L28-158
evalplus (`26d6d00b`)	n/a	Chen estimator applied to caller-supplied samples; independence not checked	n/a	evalplus/eval/__init__.py L58-84 (estimate_pass_at_k, comment 'unbiased estimator from https://github.com/openai/human-eval'); evaluate.py L330-349 (pass_at_k computation on base+plus); provider/hf.py L62-87 (num_return_sequences); codegen.py L12-20 (no seed param in signature)
lm-evaluation-harness (`b954108c`)	n/a	Chen estimator applied to caller-supplied samples; independence not checked (for humaneval_64: repeats=64, do_sample=true); pass@1 only (default humaneval.yaml / mbpp.yaml)	n/a	lm_eval/tasks/humaneval/utils.py L1-39 (pass_at_k delegates to hf_evaluate 'code_eval'); tasks/humaneval/humaneval.yaml (repeats:1, do_sample:false, k:[1]); tasks/humaneval/humaneval_64.yaml (repeats:64, do_sample:true, temperature:0.2, k:[2,8,16,32,64]); tasks/mbpp/utils.py L1-29 (pass_at_1 only, no k>=2 variant found in tree); lm_eval/api/model.py L275-350 (CachingLM); evaluator.py L81-84,196-214,409-416

D.3 Corpora

This subsection describes each corpus: where it comes from, the model and verifier, and what a run and an attempt are.

Headline design: shared-prefix continuations. Generated by us with a 7B instruction-tuned open-weight model, Qwen2.5-7B-Instruct (Qwen Team et al., 2025), in bfloat16 on 200 grade-school arithmetic questions from GSM8K (Cobbe et al., 2021): for each question 16 independently sampled reasoning prefixes of 48 tokens, each continued 8 times, with an exact-match verifier against the reference answer; a prefix is resampled if it already contains the answer or ends the generation (no early stop). A run is a prefix, an attempt is a continuation.

Reasoning tree with step labels. A public corpus of human-labelled reasoning trees on competition mathematics, PRM800K (Lightman et al., 2024), released under the MIT licence at a pinned revision. A run is a labelled node (prefix), an attempt is one of the sibling completions proposed at that node, and success is the human label "correct"; the first completion at each node is the pre-generated step and is dropped, only nodes with the full sibling count are kept, and duplicate prefixes within a question are removed. Three filters on the number R_q of sibling groups of a question set the question counts: the 8,061 questions above carry $R_q \geq 2$ and enter the $k = 2$ figures; 4,880 of them carry $R_q \geq 4$, the subset on which the $k = 4$ cross-run statistic exists and on which the main estimates are taken; and

2,358 carry $R_q \geq 8$, so that a held-out half of $R_q/2$ groups holds the four groups the first-attempt reference needs at $k = 4$ (Table 45).

Arms A and B. Two further generations of our own, laid out like the headline design but with $R = 32$ prefixes of 48 tokens and $n = 32$ continuations each. Arm A answers the same 200 grade-school arithmetic questions with a 1.5B instruction-tuned model of the same family, Qwen2.5-1.5B-Instruct, at the same 512-token limit; 194 questions are analysed, the other 6 being those for which 32 admissible prefixes could not be collected. Arm B uses MATH-500, the 500-problem held-out subset of a competition mathematics dataset (Hendrycks et al., 2021) selected uniformly at random from its test split by a process-supervision study (Lightman et al., 2024), at a pinned revision. It answers the 64 of those problems on which an independent pilot sample of 128 unshared chains scored between 0.05 and 0.95, with the 7B model and a 1024-token limit. Its answer extraction and normalisation are those of the pilot pipeline for that corpus; they differ from the arithmetic ones because the reference answers are LaTeX expressions. A truncated generation is scored as a failure in both arms and the truncation rate is reported for both. A run is a prefix and an attempt is a continuation throughout. Arm C is Arm A with no admissibility condition on the prefix (Table 19).

Arm F. A supplementary batch on the headline design's model: 128 fresh prefixes of $n = 8$ continuations each on the 62 headline questions whose within-run correlation is defined, drawn by the headline design's procedure with its prompt, sampling parameters and verifier. It serves as a fresh reference for the headline intervals (Appendix B.2; Supplement §S2.1).

Self-debugging loop. A further generation of our own: a Reflexion-style loop (Shinn et al., 2023) writes its own tests and resubmits while they fail, up to $k_{\max} = 4$ attempts, on the 164 Python problems of HumanEval (Chen et al., 2021), with $R = 16$ fresh-seed runs per problem on Qwen2.5-7B-Instruct and on Qwen2.5-1.5B-Instruct. A run is one execution of the loop and an attempt is one submitted program, scored against the hidden tests; the shared state of a run is the internal test set drawn once at its start, and attempts are ordered (Appendix B.6).

Tree-search scaffold. A further generation of our own, on the 194 arithmetic questions of Arm A in their recorded order, with the same 7B model in bfloat16 and with the prompting, sampling parameters, answer extraction and exact-match verifier imported unchanged from that pipeline. A run is one breadth-first search of depth two and branching two: two first steps are sampled, two continuations of each, a value prompt scores the four depth-two nodes greedily at eight tokens and prunes to two survivors, and each survivor is completed twice, so a run yields 4 leaves and the 512-token limit of a leaf chain is split 48, 48 and 416 across the three segments. Nothing about the reference answer enters generation or pruning, no prefix is imposed and no candidate is resampled. A run is a search and an attempt is a leaf, numbered by completion order; each generation call is seeded individually and no cache is carried between calls or between runs.

Agent corpus. A public corpus of software-engineering agent trajectories with run-level outcomes only; it is not one of the eight designs of §6 and is described in Supplement §S5.

D.4 What the supplementary material contains

The supplementary archive carries `supplement.pdf`, a technical report built from the same sources as this paper: the robustness checks, full grids and the agent corpus that this appendix points to as Supplement §§S1–S9. The supplementary archive carries, under `repro/`, the files needed to rebuild Table 13 and Tables 22 and 23 without running a model, and the replay driver that does it; everything in it is derived mechanically from the runs, with run identifiers replaced by neutral labels. `ordered-outcomes.csv` holds one row per question per run: *the ordered binary outcome of every attempt of that run*, in the order the run produced them, for the $R = 32$ by $n = 32$ arm and for the $R = 128$ by $n = 8$ batch drawn afresh on the same questions. The order is what the leading-k event is defined by, and a run's success count does not determine it: the eight-attempt outcomes 10000000 and 00000001 share a count and differ in their first-four-attempt success. `coverage-splits.csv` records the split assignment behind Table 22: which runs formed the estimation half in each of the 50 repetitions of each question, the remaining runs being the reference half. `replay-inputs.json` carries the design constants, the fixed root seeds that fix every bootstrap and simulation draw, and the recorded value of every cell the replay rebuilds. `replay_tables.py` then rebuilds those cells from those files and the estimator implementation, prints each beside its recorded value and exits non-zero on any difference: on the released bundle it reproduces every cell of Tables 22 and 23 exactly from the

ordered outcomes, splits and seeds, interval endpoints, containment averages, widths and question-level bootstrap endpoints alike, and the numeric rows of Table 13 from the per-question readings; that table's i.i.d.-control row, its two containment rows and its direction-rule row are not replayed. The comparison is exact (tolerance zero; the arithmetic is deterministic given the seeds). The fitted-law column of Table 22 is rebuilt too, from laws fitted to the same arm and the fixed simulation seed, on the simulated question set that is its denominator and not on the real questions. `table2_risk.py` recomputes the low-bias cost table (Table 4) by exact enumeration over the product grid of per-run success counts, at the support endpoints and mixing-weight grid `replay-inputs.json` carries at full precision rather than at the four decimals the table prints. It is a separate implementation rather than the same code run again: it does not import `passk_core.py`, but builds both designs' estimators from their definitions, so agreement with the recorded cells is a second reading of the same quantities. Agreement is to an absolute tolerance of 10^{-9}. `per-run-results.csv` holds one row per question per run: the number of runs, the attempts per run, the success count of every individual run, the within-run correlation estimate, and the per-question readings at each budget. `prompt.md` reproduces the instruction, both few-shot exemplars, the chat-template construction and the answer-extraction and normalisation rules verbatim from the generation code stored inside the run directories. `settings.tsv` lists each run's model, design and sampling parameters as its own generation metadata recorded them. `filtering.md` and `prefix-rejects-by-question.csv` carry the question and prefix filtering record: how many prefix draws were discarded and redrawn, per question and per rejection reason, and which questions were excluded and why. `fresh-batches.md` says which run is the parent design and which the independently generated reference batch, with the counters that show nothing was copied between them. `table-manifest.md` gives, for every printed cell of those tables, the key path inside the run's JSON it was selected from; it is the verbatim output of each generator's own manifest flag. `passk_core.py` is the estimator implementation the runs themselves used, which `replay_tables.py` imports rather than restates — `table2_risk.py` does not import it — and `run-provenance.md` maps every table in the paper to the run behind it, under those labels. Binary outcomes settle the statistical formulas but not whether the verifier, the stopping rule, the state reset and the scaffold match this description, so `scaffold/` carries the generation and scoring programs of the two generated scaffolds — the self-debugging loop and the tree search — and the three local helper modules they import, as the runs executed them, with comments, docstrings and report/message strings translated into English, run identifiers relabelled and filesystem paths replaced (the supplement's `README.md` gives the token-level check, the recorded hashes and a map of which components are included and which are referenced but absent); `ordered-outcomes-scaffold.csv` and `ordered-outcomes-treesearch.csv` give every attempt and every leaf of those arms in order, each attempt carrying the run's own verifier decision beside the held-out one; and `trajectories/` holds ten complete self-debugging runs — the tests the run wrote for itself, then every attempt's submitted code, the feedback returned and both readings — selected by the stated rule of taking the first ten in the file's own order. The bundle needs Python 3.12 or later, because the bootstrap draws use `random.binomialvariate`; only `table2_risk.py` needs a third-party package (NumPy). The raw model continuations are not included: they are large, and no estimator in this paper reads the text of a continuation — each reads the ordered pass/fail outcomes above, or the counts derived from them. Nothing from the observational agent corpus is in this archive, and no cell replayed here is computed from it: the extraction code and derived per-run success table named in Supplement §S5 belong to the code release that accompanies publication, which cannot be shipped anonymously.

E Related work in detail

Table 42 draws the boundary paper by paper; this section states what we build on. *The two estimands and the sign of their gap.* Tran-Truong and Le (2026, Thm. 4) state the inequality for trials sharing a latent state, with its equality case, and Bay and Yearick (2026, Prop. 3) restate it as a correlation tax, bridge it to the within-run correlation and separate the between-problem difficulty spread from the within-problem one. We build on this distinction; our finite central-moment expansion (Corollary A.5) makes its higher-order terms explicit. *The name* succ@k. Liu et al. (2025) report, under the name Succ@k, the fraction of problems solved within k feedback turns, next to a parallel-sample pass@k of a different model, and read it as the probability of success within k attempts; that is the event of (2). We build on it; our contribution is its treatment as an estimand: which observations identify it, its uniqueness (Theorem 3.5) and run-level inference for it. *The estimability criterion is classical.* A functional of a law is unbiasedly estimable from m independent draws exactly when

it is a polynomial of degree at most m in the law (Halmos, 1946), restated for black-box generative losses by Kimpara et al. (2023, Lem. A.3.1–A.4.1), who attribute it to Kolmogorov and to Lehmann and Casella. We inherit it whole; Theorem 4.1 contributes the identification of the draw it counts as a complete execution observed through any kernel fixed in advance of the unknown mixing law, and Theorem A.6 the approximation characterisation of the residual bias below the threshold. *The variance floor builds on prior work in the same way.* The hierarchical decomposition of a mean over clusters of repeated draws already contains the $k = 1$ case of Theorem A.20 (Zhang et al., 2026, Lem. 2.1); the design effect $n/[1 + (n - 1)\rho]$ (Kerry and Bland, 1998; Bay and Yearick, 2026) is the same saturation reached from an estimator's variance rather than from estimability; and run-level repetition has already been urged on empirical grounds, paired with an intraclass correlation, made the replicated unit of computer-use benchmarking and given a repaired operational definition (Bjarnason et al., 2026; Mustahsan et al., 2025; D'Oro et al., 2026; Jiang et al., 2026), row by row in Table 42. *Allocation across questions.* Kazdan et al. (2025, Thm. 1, App. D) give the asymptotically variance-optimal per-question budget $b_q^\star \propto \sqrt{p_q(1 - p_q)^{2k-1}}$ for independent draws; §5.4 prices a different trade, between fresh executions and attempts sharing one latent state within a question, at the exact finite-n variance of the stated estimator. What remains ours is the list of §1, and Table 42 sets it against each neighbour row by row. Anytime-valid evaluation certifies a plurality answer under an i.i.d. assumption on attempts (Ota et al., 2026; Cordero-Encinar and Duncan, 2025), while Hariri et al. (2026b); Brown et al. (2024) keep independence; the confidence-sequence construction is standard and its role here is a run-level evaluation protocol.

GDAG (Mohri et al., 2026, §4.4, §5.1–5.4) already combines binomial polynomial estimability with minimax uniform approximation for a gradient-weighted $p \log p$ target. The distinction here is the $R \geq k$ threshold for general calibrated run-observation kernels and the attained bias on a specified two-point family, alongside the restricted budget-k estimand. PACE (Shawn, 2026, §4) already discards ties and bets on discordant pairs for anytime-valid, per-candidate acceptance under a conditional null. Our standard Beta-mixture confidence sequence instead quantifies complete-run budget-k success and supports paired comparison; the discordant-pair reduction itself is not a new contribution.

Table 42: Each neighbouring work differs from this paper in the estimand, the estimator class, the sampling regime or the assumption that carries the argument; the table states the relation in one phrase per work. The precise reason for each row is in Table S29 of the supplementary report.

Work	Relation
Fair scores for finite dependent ensembles in forecast verification (Ferro, 2014)	same technique, orthogonal problem
Asymptotic bias of majority-vote reward estimation under exchangeable latents (Du et al., 2026)	related mechanism, different target
Unbiased estimation of p^k from n Bernoulli trials (Lehmann and Casella, 1998)	strict special case
The same count restated for policy optimisation (Walder and Karkhanis, 2025)	same classical count, different use
Polynomial characterisation of unbiasedly estimable functionals from i.i.d. draws (Halmos, 1946; Kimpara et al., 2023)	same technique, different draw
What a fixed rollout depth identifies (Singh and Singh, 2026)	concurrent; identification across tasks
GDAG: polynomial estimation and minimax bias (Mohri et al., 2026)	same approximation route, different target and observation
Run-level repetition in agent and computer-use evaluations (Bjarnason et al., 2026; Mustahsan et al., 2025; D'Oro et al., 2026)	prior on the unit; empirical and operational grounds
Dyadic subsampling for quasi-Monte-Carlo samplers of language models (Li et al., 2026)	compatible; an existence-side example

continued on the next page

Table 42, continued.

Work	Relation
Anytime-valid certification of self-consistency (Ota et al., 2026; Cordero-Encinar and Duncan, 2025)	orthogonal target
PACE: anytime-valid candidate acceptance (Shawn, 2026)	same paired reduction, different evaluation target
Anytime-valid allocation of the question-sampling budget (Zhou et al., 2026)	partially adjacent
Question-level standard errors for repeated decodes (Miller, 2024)	different unit, different target
Hierarchical variance of a benchmark mean under repeated generations (Zhang et al., 2026)	prior on the variance floor; nothing on estimability
Operational redefinition of the repetition count (Jiang et al., 2026)	prior on the unit; silent on which estimand
Extrapolating the pass@k curve across a benchmark, and optimal allocation across problems (Kazdan et al., 2025)	prior on allocation; a different failure of the same estimator
Ceilings on test-time scaling under correlated attempts (Bay and Yearick, 2026)	nearest antecedent on the estimand; orthogonal on estimability
Effective-count correction of the coverage curve (Levi, 2024)	earlier member of the effective-count family; no estimability content
Markov-chain reliability of agent traces (Tran-Truong and Le, 2026)	the earlier of the two explicit statements of the estimand gap cited here; nothing on the protocol side
Multi-turn reasoning with unary feedback (Liu et al., 2025)	prior use of the name; a two-model comparison, no identification or intervals
Repeated sampling and adaptive stopping for test-time scaling (Wang et al., 2023; Aggarwal et al., 2023)	background; different target (the majority answer)

F Extended arguments

Each section below expands a passage the body states in one or two sentences, and the body points to it at that place.

F.1 The degree argument behind the threshold

Plainly: below k fresh runs the independent-attempt number is not merely hard to estimate; nothing is unbiased for it. The mechanism is a degree count: along the two-point family π_α the law of one run seen through K is affine in α, so the joint law of R independent runs, and with it every estimator's expectation, is a polynomial in α of degree at most R, one degree per independent run, while the target $1 - (1 - \theta_0 - \alpha\Delta)^k$ has degree exactly k. When each run emits a single Bernoulli trial this is the classical fact that p^k is unbiasedly estimable from n i.i.d. trials if and only if $n \geq k$ (Lehmann and Casella, 1998; restated in Kimpara et al., 2023, Lem. A.4.1), the Bernoulli case of the characterisation of unbiasedly estimable functionals as polynomials of degree at most the sample size (Halmos, 1946; restated in Kimpara et al., 2023, Lem. A.3.1): the necessity half is that classical argument with the run in place of the draw, and Theorem 4.1(b) is Le Cam's two-point argument. The calibrated-kernel sufficiency step extracts one Bernoulli bit per run and applies the classical distinct-run product construction. Thus the degree criterion and that estimator construction are inherited tools; the specialization here identifies the independent run as their sampling unit, alongside the finite-design and approximation results compared in §2.

Table 43: Of the four numbers "pass@2" can denote for the two-attempt procedure of Example F.1 — the three i.i.d. readings and succ@2 — only succ@2 is a property of the procedure. The subset-averaged success probability (Proposition 3.3) is shown for reference; with $n = k = 2$ its only subset is the pair of attempts, so here it equals succ@2. In that procedure the second attempt is suppressed whenever the first succeeds; columns: the reading, the single-attempt rate it substitutes into (1), and the value. Exact rationals, no experimental input.

Reading	Substituted rate	Value
i.i.d. counterfactual at the first-attempt rate	$p = \mathbb{P}(X_1 = 1) = 0.2$	0.360
i.i.d. counterfactual at the pooled marginal	$p = 0.3$	0.510
i.i.d. counterfactual at the second-attempt rate	$p = \mathbb{P}(X_2 = 1) = 0.4$	0.640
Subset-averaged success probability	—	0.600
succ@2, what the procedure does	—	**0.600**

F.2 The distinction is not cosmetic

Without exchangeability there is no single "single-attempt success rate" to substitute into (1): the first attempt, the second and the pooled average have different marginals, and with two attempts the three i.i.d. readings already disagree, none of them the probability that the procedure succeeds (Example F.1; values in Table 43).

Example F.1 (Four readings of one procedure)**.** Let $n = k = 2$ with $X_1 \sim \text{Bern}(0.2)$, $X_2 \mid X_1 = 0 \sim \text{Bern}(0.5)$ and $X_2 \mid X_1 = 1 \equiv 0$. The procedure succeeds within two attempts with probability succ@2 = 0.600. Substituting a single-attempt rate into (1) gives three other numbers: 0.360 at the first-attempt rate 0.2, 0.510 at the pooled rate 0.3, and 0.640 at the second-attempt rate 0.4 (Table 43).

F.3 Effective attempts against the estimability threshold

On Arm A (§6.3) the effective attempt count and the threshold $R \geq k$ give opposite advice (Table 36): at an average within-run correlation $\bar{\rho} = 0.6838$, three runs of 32 continuations carry 4.325 effective attempts against 4.000 for four single-attempt runs, so by the count the first design is the better instrument for pass@4. If uniform unbiasedness over the stated class is required, the threshold excludes the first design, and a held-out reference agrees: scored against the other half of each question's runs, the three-run design's pooled estimate has mean error −0.0487 and its percentile-bootstrap interval over questions contains the reference 58% of the time, against +0.0006 and 99% for four single-attempt runs, whose interval is the wider one (Table S13); the difference is centring, not width. The effective count is right about precision; what it cannot see is that unbiased estimability of the fixed target begins at $R = k$.

F.4 The run-level confidence sequence

This subsection defines the run-level confidence sequence, the two benchmark sampling schemes and the paired comparison, and what a benchmark-level interval costs in runs. For a Beta(a, b) mixture e-process with $A_t = \sum_{i \leq t} Y_i$,

$$M_t(s) \;:=\; \frac{B(a + A_t,\; b + t - A_t)}{B(a, b)\; s^{A_t}(1 - s)^{t - A_t}}, \qquad \textsc{run}\text{CS}_t(\delta) \;:=\; \{s \in [0, 1] : M_t(s) < 1/\delta\}, \tag{17}$$

Ville's inequality (Ville, 1939) gives $\mathbb{P}(\forall t : s_q \in \textsc{run}\text{CS}_t(\delta)) \geq 1 - \delta$ (Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Ramdas et al., 2023), at any data-dependent stopping time; we use the Krichevsky–Trofimov choice $a = b = \frac{1}{2}$ (see Ryu and Wornell, 2024, §3.1). At the runs one question can afford the sequence is wide — a width of 0.10 at $s = 0.5$ needs 1,359 runs (Table 44) — so it belongs at the benchmark level, where $\bar{s} := Q^{-1} \sum_q s_q$ is the target. Two sampling schemes reach it with no multiplicity correction: sampling one question per step (scheme W) makes $Y_t \sim \text{Bern}(\bar{s})$ i.i.d., so (17) applies verbatim; running every question once per round (scheme S) makes the round means i.i.d. with mean $\bar{s}$, for a bounded-mean sequence (Waudby-Smith and Ramdas, 2024). A paired comparison reduces to the same stream on the discordant pairs, and to certify $\bar{s} \geq \tau$ one stops at the first t with $\inf \textsc{run}\text{CS}_t(\delta) \geq \tau$, a false declaration then having probability at most δ under any stopping rule (measured in Table S27(b)).

The efficiency ratio of the two benchmark sampling schemes. Scheme W samples one question per step and records its run indicator; scheme S runs every question once per round and records the round mean (§F.4). The variance of one round of scheme S is $Q^{-2}\sum_q s_q(1-s_q)$, against $\bar{s}(1-\bar{s})/Q$ for the mean of Q steps of scheme W, the same number of runs, so the efficiency ratio

$$e_{\mathrm{WS}} \;=\; \frac{\sum_q s_q(1-s_q)}{Q\,\bar{s}(1-\bar{s})} \tag{18}$$

is exactly one minus the share of variance carried by question-difficulty heterogeneity; it is measured on the agent corpus in Supplement §S5.

For a comparison of systems A and B, run both on the same drawn question and record $D_t := Y_t^A - Y_t^B \in \{-1, 0, 1\}$, i.i.d. with mean $\Delta = \bar{s}_A - \bar{s}_B$. We certify through the *discordant pairs*: updating only on steps with $D_t \neq 0$ and writing $p_+ := \mathbb{P}(D = +1 \mid D \neq 0)$, we have $\Delta > 0 \iff p_+ > \frac{1}{2}$, so the Beta$(\frac{1}{2}, \frac{1}{2})$ mixture e-process of (17) applies to the Bernoulli subsequence $\mathbf{1}\{D_t = +1\}$, and optional skipping preserves time-uniform validity. Pairing removes the common variance due to question difficulty; the discordant-pair reduction additionally removes the steps in which both systems fail, which carry no information about the sign of Δ.

A bounded-difference sequence on $(D_t+1)/2$ pays the worst-case variance $t/4$ regardless of the data, and in a paired comparison most steps carry $D_t = 0$ — both systems fail on the same hard item — so the bound is wide exactly where the information is thin; restricting the sequence to the discordant steps, where all of the evidence about the sign of Δ lives, is not a repair of that bound but the right unit for this question.

Multiple budgets. A procedure satisfying Assumption A.2 yields all $Y^{(k)}$ for $k \le k_{\max}$ from one run at budget $k_{\max}$; simultaneous validity over a grid of m budgets costs a δ/m union bound. A procedure that does not satisfy A.2 must be run separately at each budget: by Theorem 3.5 the estimand is indexed by the budget the procedure is actually permitted.

F.5 Valid intervals for pass@k^{iid}

Theorem 4.1 rules out a uniformly unbiased point estimator of pass@k^{iid} below $R = k$. It does not rule out inference about pass@k^{iid}, and the separation is worth making with the same data. Take the first attempt of each run, $B_r := X_1^{(r)}$. Definition A.1 makes B_r a measurable function of the observation with $\mathbb{P}(B_r = 1 \mid \theta_r) = \theta_r$, and runs are independent, so $B_1, \ldots, B_R$ are i.i.d. Bern(p) with $p = \mathbb{E}_\pi[\theta]$. If $[L, U]$ covers p with probability at least $1 - \delta$ — at a fixed sample size, or at every t simultaneously for a sequence — then, because $g_k(p) = 1 - (1-p)^k$ is strictly increasing on $[0, 1]$, $\{p \in [L, U]\} = \{g_k(p) \in [g_k(L), g_k(U)]\}$ as events, so $[g_k(L), g_k(U)]$ covers pass@k^{iid} at the same level, at any data-dependent stopping time when the input is a sequence. Nothing in this asks for $R \ge k$; what the threshold costs below k is width. The transform is applied per question: g_k is strictly concave for $k \ge 2$, so $Q^{-1}\sum_q g_k(p_q) \le g_k(Q^{-1}\sum_q p_q)$, strictly unless all p_q coincide, and an interval built on a question-averaged p is not an interval for the average of the per-question targets.

A second construction uses all n attempts. Under Definition 3.1 the run mean $\bar{X}^{(r)} := n^{-1}\sum_i X_i^{(r)}$ lies in $[0, 1]$ and has $\mathbb{E}[\bar{X}^{(r)}] = \mathbb{E}[\theta] = p$, so a bounded-mean sequence (Waudby-Smith and Ramdas, 2024) on $\bar{X}^{(1)}, \bar{X}^{(2)}, \ldots$ transforms the same way. The sequence itself needs only independent runs; what ties its target to p is that every attempt position has success rate p, which exchangeability inside the run guarantees: where the attempts are ordered, $\mathbb{E}[\bar{X}^{(r)}]$ is the position-averaged attempt rate and need not equal μ_1, and the transformed interval is then an interval for something else. Both facts are visible in the tables below.

Tables S30 and S31 score the two constructions, and a fixed-sample Clopper–Pearson companion, on every design of §6 that records ordered attempt outcomes, beside the three succ@k intervals of the other interval tables of this appendix. Every interval sits on the estimation half of a split, and the three succ@k constructions reproduce those tables' numbers. Table 48 reads Arm A against a 128-run fresh batch as well as against the held-out half, once as succ@k and once as pass@k^{iid}: the run-level succ@k intervals and the two first-attempt constructions each contain the reference for their own estimand far more often than the other one, which is the separation stated as a measurement; the

Table 44: At the run counts one question can afford a run-level interval is wide: a width of 0.10 at $s = 0.5$ needs 1,359 runs for run-level runCS. How many runs a target precision costs, at $\delta = 0.05$ (computed; no data enter this table). For a run-level success probability s and a target width w, the columns give the smallest number of runs R at which run-level runCS of (17) and the run-level Clopper–Pearson interval have width at most w, together with the width attained there; the observed success count at R runs is taken to be the nearest integer to sR, and the reported R is the smallest one from which the width stays below the target for every larger R in the searched range, since that rounding makes the width non-monotone by a few runs. Every count here is the cost of *the two constructions named in the columns* — the Beta$(1/2, 1/2)$-mixture Bernoulli sequence of (17) and the Clopper–Pearson interval on the same run indicators — at the stated width. It is not a sample-complexity lower bound over run-level procedures: a different observation, a different prior or a different level of time-uniformity buys a different curve, and Table 49 measures two such alternatives. What Proposition A.9 adds, and what these counts illustrate rather than prove, is that the *variance* available to any unbiased estimator is governed by the number of runs and not by the number of attempts inside them, so no amount of resampling within a run moves it. The table also sets the scale of the width reported in Table 10, where eight runs give a run-level interval of width 0.48: at $s = 0.5$ this sequence reaches a width of 0.10 only at 1,359 runs and 0.05 at 6,053. Two consequences are worth stating. Anytime validity is not free: across these cells the confidence sequence needs between 2.8 and 3.8 times the runs of the fixed-sample interval, which is the price of being allowed to stop whenever one likes. And a benchmark reported at run level is a different budget conversation from one reported at attempt level: with n attempts per run the same precision costs n times the attempts, which is why §5.4 sends the total cost into more runs rather than into deeper ones.

s	width w	runs, runCS	attained	runs, Clopper–Pearson	attained
0.3	0.10	1,126	0.1000	341	0.0998
0.5		1,359	0.1000	402	0.0999
0.7		1,125	0.1000	341	0.0998
0.9		451	0.0999	159	0.0996
0.3	0.05	5,023	0.0500	1,328	0.0500
0.5		6,053	0.0500	1,574	0.0500
0.7		5,023	0.0500	1,328	0.0500
0.9		2,027	0.0500	592	0.0499

two wider run-mean constructions contain both. Table 50 keeps the three quantities apart at $k = 4$; its last column is the pooled excess $\widehat{C}_{\text{pooled}} - \overline{Y}$ that §6 reports for every design.

Replacing the binary run indicator Y_r by the $[0, 1]$-valued summary $Z_r = \widehat{C}_{n,k}(X^{(r)})$ does not shorten the interval here. On the headline design the run-level Bernoulli sequence on Y_r has mean width 0.477 (the transformed first-attempt sequence, which targets pass@k^{iid} rather than succ@k, has 0.150); the same Beta-mixture sequence applied to $Z_r = \widehat{C}_{n,k}(X^{(r)})$, whose average is the UMVUE of succ@k, has 0.479, and on Arm A 0.505 against 0.502 (Table 49). The mixture construction is valid for $[0, 1]$-valued observations because $(q/s)^Z((1-q)/(1-s))^{1-Z}$ is convex in Z and so bounded by its chord, whose conditional expectation is one; the wider empirical-Bernstein bounded-mean sequence (0.950 on the headline design) is kept in the table to show that its extra width is the construction's, not the summary's. Where $n = k$ the comparison is exact, since $Z_r = Y_r$ identically.

Two further readings of the same splits are recorded here as measurements. The first prices unbiasedness above the threshold. Table S32 of the supplementary report puts the biased plug-in $g_k(\bar{B})$, built from the fraction $\bar{B}$ of first attempts that succeed on the estimation half, beside the unbiased $\widehat{U}_k$ on that same half, both scored against the held-out reference of that design. On the headline design at $k = 4$ the plug-in carries bias -0.0061 against -0.0002 and mean squared error 0.0073 against 0.0034, a ratio of 2.126; on the Arm A at $k = 4$ the biases are -0.0161 and -0.0009, the errors 0.0204 and 0.0139, and the ratio 1.469. This comparison changes the attempts used (all nR against the R first attempts) as well as the estimator; Table 45 separates the two. The second reading prices width against containment on one scale, since containment alone favours a wide interval. Table 47 reports the Winkler interval score at $\delta = 0.05$ and $k = 4$ for every design (every budget: Table S33 in the supplementary report). On the headline design the attempt-level bootstrap scores 0.791, the run-level Clopper–Pearson 0.462, runCS 0.502, and the transformed first-attempt sequence, against its own pass@k^{iid} reference, 0.159; on Arm A the same four are 6.003, 0.652, 0.578 and 0.440: once width is charged, runCS still scores below the attempt-level bootstrap on both blocks.

Three further readings of the same random half-splits back §5.2 and §5.3. Table 45 compares four estimators of pass@k^{iid} at matched information — the plug-in and the unbiased statistic, each on all attempts and on first attempts only — so that the comparison of supplementary Table S32 is separated into its two parts: per question the all-attempt pair differ little, while on the shared-prefix designs and the reasoning tree the plug-in's negative bias decides the benchmark-level error. Table 46 adds the run-clustered bootstrap around $\widehat{C}_{\text{pooled}}$ to the three intervals of supplementary Table S33, and Table 47 lists the widths and interval scores behind Table 1, including the run-level Clopper–Pearson interval and the transformed first-attempt sequence.

Table 45: Per question the plug-in and the unbiased statistic on all attempts differ little in mean squared error. At the benchmark level the plug-in's bias is negative and its error the larger on every shared-prefix and reasoning-tree row; on the tree search both reverse. Four estimators of pass@k^{iid} on the same estimation halves of the same random half-splits (Tables S30 and S31), scored against the same held-out reference: the plug-in $g_k(\bar{X})$ on all attempts (AP) and the cross-run $\widehat{U}_k$ (AU), and the plug-in and unbiased estimators on first attempts only (FP, FU). Per-question MSE is averaged over splits and then over questions with equal weight; ΔMSE is AP minus AU with a 95% question-bootstrap interval. Benchmark bias and RMS take, for each split index, the question-averaged error, and then its mean and root mean square over split indices. The reference is the held-out $\widehat{U}_k$ on the exchangeable designs and the held-out first-attempt form $\widehat{U}_k^1$ on the ordered ones (†); on Arm A the second row scores the same estimates against the 128-run fresh $\widehat{U}_4$. $\widehat{U}_k$ is undefined on the self-debugging designs, whose run lengths are irregular. Q counts the questions scored; on the reasoning tree they are the 2,358 questions with at least 8 sibling groups, whose held-out half then holds the 4 groups the first-attempt reference needs at $k = 4$. References carry noise of their own, so errors are not distances to a known parameter.

		per question			benchmark level				first attempts	
Design	Q	MSE AP	MSE AU	ΔMSE [95% CI]	bias AP	bias AU	RMS AP	RMS AU	RMS FP	RMS FU
Shared prefix, 7B, $k = 4$	200	0.0036	0.0034	+0.000170 [+0.000007, +0.000374]	−0.0033	−0.0002	0.0054	0.0043	0.0084	0.0062
Shared prefix, 1.5B, Arm A, $k = 4$	194	0.0139	0.0139	−0.000014 [−0.000215, +0.000198]	−0.0101	−0.0009	0.0131	0.0084	0.0187	0.0097
fresh 128-run reference	194	0.0095	0.0094	+0.000096 [−0.000328, +0.000511]	−0.0111	−0.0019	0.0119	0.0047	0.0183	0.0070
Shared prefix, 1.5B, Arm A, $k = 16$	194	0.0212	0.0230	−0.001839 [−0.005016, +0.000509]	−0.0164	−0.0027	0.0189	0.0103	0.0382	0.0171
Shared prefix, 7B, Arm B, $k = 4$	64	0.0015	0.0015	+0.000033 [+0.000002, +0.000076]	−0.0022	−0.0000	0.0054	0.0049	0.0178	0.0104
Shared prefix, 7B, Arm B, $k = 16$	64	0.0008	0.0008	+0.000026 [+0.000005, +0.000056]	−0.0020	−0.0009	0.0038	0.0034	0.0131	0.0129
Shared prefix, 1.5B, Arm C, $k = 4$	200	0.0150	0.0149	+0.000031 [−0.000184, +0.000247]	−0.0102	−0.0000	0.0126	0.0076	0.0165	0.0085
Shared prefix, 1.5B, Arm C, $k = 16$	200	0.0158	0.0157	+0.000113 [−0.000863, +0.001000]	−0.0136	+0.0008	0.0162	0.0094	0.0331	0.0136
Reasoning tree (public), $k = 4^\dagger$	2358	0.1560	0.1558	+0.000193 [−0.001083, +0.001444]	−0.0631	−0.0079	0.0635	0.0110	0.1107	0.0099
Self-debugging loop, 7B, $k = 2^\dagger$	164	0.0192	n/a	n/a	−0.0210	n/a	0.0220	n/a	0.0112	0.0082
Self-debugging loop, 7B, $k = 4^\dagger$	164	0.0180	n/a	n/a	+0.0059	n/a	0.0098	n/a	0.0108	0.0083
Self-debugging loop, 1.5B, $k = 2^\dagger$	164	0.0501	n/a	n/a	−0.0787	n/a	0.0794	n/a	0.0197	0.0115
Self-debugging loop, 1.5B, $k = 4^\dagger$	164	0.0643	n/a	n/a	−0.0591	n/a	0.0613	n/a	0.0323	0.0181
Tree search, 7B, $k = 4^\dagger$	194	0.0099	0.0102	−0.000314 [−0.000899, +0.000201]	+0.0040	+0.0078	0.0069	0.0096	0.0106	0.0072

Table 46: Resampling complete runs around the pooled estimate does not reach the containment of the run-level intervals on any design: the centre is still the pooled number. Resampling complete runs instead of attempts, around the pooled estimate, at $k = 4$ on the same random half-splits of every design as the other interval tables. The two run-clustered rows resample the R' runs of the estimation half with replacement and pool each resample again; the bootstrap law of $\widehat{C}^*_{\text{pooled}}$ is computed exactly (no Monte-Carlo error), and the interval is either its 2.5% and 97.5% quantiles or $\widehat{C}_{\text{pooled}} \pm 1.96$ times its standard deviation, clipped to $[0, 1]$. The other three rows are the intervals of Table S33, recomputed (identical). Containment is of the held-out $\overline{Y}_4$ (and, on Arm A, of the 128-run fresh $\overline{Y}_4$); zero width is the fraction of intervals whose two endpoints coincide; the Winkler score at $\delta = 0.05$ charges width and misses on one scale (lower is better).

	held-out $\overline{Y}_4$				fresh $\overline{Y}_4$
Interval	containment	mean width	zero width	Winkler	containment
Shared prefix, 7B (headline)					
attempt-level bootstrap	0.864	0.028	0.769	0.791	—
run-clustered bootstrap, percentile	0.902	0.048	0.769	0.493	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.911	0.043	0.734	0.606	—
run-level RUNCS	0.966	0.477	0.000	0.502	—
run-level Clopper–Pearson	0.960	0.402	0.000	0.462	—
Shared prefix, 1.5B, Arm A					
attempt-level bootstrap	0.146	0.052	0.069	6.003	0.104
run-clustered bootstrap, percentile	0.416	0.203	0.069	3.363	0.327
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.396	0.191	0.102	4.009	0.270
run-level RUNCS	0.892	0.502	0.000	0.578	0.990
run-level Clopper–Pearson	0.838	0.386	0.000	0.652	0.952
Shared prefix, 7B, Arm B					
attempt-level bootstrap	0.208	0.029	0.048	1.614	—
run-clustered bootstrap, percentile	0.310	0.063	0.048	1.068	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.659	0.061	0.090	1.186	—
run-level RUNCS	0.922	0.374	0.000	0.402	—
run-level Clopper–Pearson	0.902	0.288	0.000	0.395	—
Shared prefix, 1.5B, Arm C					
attempt-level bootstrap	0.190	0.053	0.113	6.399	—
run-clustered bootstrap, percentile	0.446	0.215	0.113	3.516	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.413	0.202	0.136	4.226	—
run-level RUNCS	0.905	0.504	0.000	0.570	—
run-level Clopper–Pearson	0.858	0.387	0.000	0.636	—
Reasoning tree (public)					
attempt-level bootstrap	0.609	0.492	0.081	4.928	—
run-clustered bootstrap, percentile	0.620	0.554	0.124	5.017	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.624	0.490	0.123	5.416	—
run-level RUNCS	0.788	0.860	0.000	1.257	—
run-level Clopper–Pearson	0.772	0.804	0.000	1.412	—
Self-debugging loop, 7B					
attempt-level bootstrap	0.785	0.078	0.558	0.745	—
run-clustered bootstrap, percentile	0.830	0.084	0.560	0.653	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.877	0.073	0.518	0.745	—
run-level RUNCS	0.925	0.482	0.000	0.505	—
run-level Clopper–Pearson	0.920	0.406	0.000	0.470	—
Self-debugging loop, 1.5B					
attempt-level bootstrap	0.670	0.248	0.274	2.011	—
run-clustered bootstrap, percentile	0.768	0.303	0.274	1.513	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.762	0.276	0.251	1.897	—
run-level RUNCS	0.887	0.589	0.000	0.681	—
run-level Clopper–Pearson	0.873	0.497	0.000	0.720	—

Table 46, continued.

Tree search, 7B					
attempt-level bootstrap	0.913	0.041	0.779	0.530	—
run-clustered bootstrap, percentile	0.931	0.054	0.779	0.379	—
run-clustered bootstrap, $\widehat{C}_{\text{pooled}} \pm z\,\text{sd}^*$	0.941	0.049	0.759	0.469	—
run-level RUNCS	0.984	0.473	0.000	0.492	—
run-level Clopper–Pearson	0.976	0.398	0.000	0.446	—

Table 47: Mean width and mean Winkler score ($\delta = 0.05$) at $k = 4$ of the four constructions of Table 1, plus the run-level Clopper–Pearson interval. The three succ@k intervals are scored against the held-out $\overline{Y}_4$, the transformed first-attempt sequence against its own pass@k^{iid} reference (held-out $\widehat{U}_4$ on the shared-prefix designs, $\widehat{U}_4^1$ on the others). Values are those of Table S33.

	mean width				Winkler score			
Design	attempt	runCS	CP	g_4(runCS$_1$)	attempt	runCS	CP	g_4(runCS$_1$)
Shared prefix, 7B (headline)	0.028	0.477	0.402	0.150	0.791	0.502	0.462	0.159
Shared prefix, 1.5B, Arm A	0.052	0.502	0.386	0.418	6.003	0.578	0.652	0.440
Shared prefix, 7B, Arm B	0.029	0.374	0.288	0.286	1.614	0.402	0.395	0.291
Shared prefix, 1.5B, Arm C	0.053	0.504	0.387	0.428	6.399	0.570	0.636	0.457
Reasoning tree (public)	0.492	0.860	0.804	0.894	4.928	1.257	1.412	1.525
Self-debugging loop, 7B	0.078	0.482	0.406	0.219	0.745	0.505	0.470	0.237
Self-debugging loop, 1.5B	0.248	0.589	0.497	0.527	2.011	0.681	0.720	0.670
Tree search, 7B	0.041	0.473	0.398	0.157	0.530	0.492	0.446	0.181

Table 48: The run-level intervals for succ@k contain the succ@k reference far more often than the pass@k^{iid} reference, and the first-attempt intervals for pass@k^{iid} the reverse. The same $k = 4$ intervals of Arm A, on the same 194 questions and the same splits, scored against four references: the held-out half and a 128-run batch drawn afresh on the same questions, each read once as succ@k (through $\overline{Y}_4$) and once as pass@k^{iid} (through the cross-run $\widehat{U}_4$). The first three constructions target succ@k, so their pass@k^{iid} columns are shown to make the mismatch visible and are not a defect of those intervals; the last four target pass@k^{iid} and their succ@k columns read the same way. No column is coverage of a known parameter. The fresh reference is the less noisy of the two, and it is where the transformed first-attempt sequence is read.

	vs succ@k reference		vs pass@k^{iid} reference		
interval	held-out	fresh	held-out	fresh	mean width
attempt-level bootstrap	0.146	0.104	0.282	0.308	0.052
run-level runCS on Y_r	0.892	0.990	0.510	0.542	0.502
run-level Clopper–Pearson on Y_r	0.838	0.952	0.416	0.448	0.386
g_k(runCS on first attempts)	0.594	0.627	0.868	0.990	0.418
g_k(Clopper–Pearson on first attempts)	0.461	0.454	0.811	0.963	0.325
g_k(bounded-mean sequence on $\bar{X}^{(r)}$)	0.901	0.925	0.999	1.000	0.628
g_k(fixed-sample emp. Bernstein on $\bar{X}^{(r)}$)	1.000	1.000	1.000	1.000	0.923

Table 49: Replacing the binary run indicator Y_r by the within-run classical estimator $Z_r = \widehat{C}_{n,k}(X^{(r)}) \in [0, 1]$, whose average is the UMVUE of succ@k under the conditions of Proposition A.14, does not shorten the interval: under the same Beta-mixture construction the two widths agree, and the extra width of the bounded-mean rows is that construction's. Headline design and Arm A at $k = 4$; same questions, splits, estimation halves and δ, every row scored against the held-out $\overline{Y}_k$. The Beta-mixture sequence on Z_r is valid for any $[0, 1]$-valued observation with the stated conditional mean. Every design and budget: Table S34 of the supplementary report.

Corpus / model	k	interval	containment	mean width	median width
Shared prefix, 7B (headline)	4	run-level runCS on Y_r	0.966	0.477	0.439
	4	run-level Clopper–Pearson on Y_r	0.960	0.402	0.369
	4	Beta-mixture sequence on Z_r	0.959	0.479	0.439
	4	bounded-mean sequence on Z_r	1.000	0.950	0.941
	4	fixed-sample emp. Bernstein on Z_r	1.000	1.000	1.000

continued on the next page

Table 49, continued.

Corpus / model	k	interval	containment	mean width	median width
Shared prefix, 1.5B, Arm A	4	run-level runCS on Y_r	0.892	0.502	0.540
	4	run-level Clopper–Pearson on Y_r	0.838	0.386	0.416
	4	Beta-mixture sequence on Z_r	0.893	0.505	0.540
	4	bounded-mean sequence on Z_r	0.999	0.716	0.709
	4	fixed-sample emp. Bernstein on Z_r	1.000	0.960	1.000

Table 50: The three quantities kept apart at $k = 4$, on the same runs, each with a 95% percentile question-bootstrap interval. Column 1, $\widehat{U}_4 - \overline{Y}$, is unbiased for the estimand gap pass@k^{iid} – succ@4. $\widehat{U}_4$ is the cross-run U-statistic of Proposition A.8 on the four shared-prefix rows, its random-sibling form on the reasoning tree (Table 18; the held-out reference for that row in Tables 7 and 45 is the first-sibling form), and its first-attempt form $\widehat{U}_4^1$ on the two self-debugging rows, whose run lengths are irregular, and on the tree search. Columns 2 and 3 read the pooled report $\widehat{C}_{\text{pooled}}$ against each estimand in turn; column 3 is the pooled excess. All three are observed differences between estimates on one corpus, not biases against a known truth: column 1 is unbiased for a difference of parameters, while columns 2 and 3 combine the pooled estimator's finite-design centring with the sampling noise of both terms. The point estimates of column 1 equal those of Table 7; the bootstrap endpoints differ slightly because the resamples use different random seeds.

Corpus / model	Q	$\widehat{U}_4 - \overline{Y}$	$\widehat{C}_{\text{pooled}} - \widehat{U}_4$	$\widehat{C}_{\text{pooled}} - \overline{Y}$
Shared prefix, 7B (headline)	200	+0.0262 [+0.0161, +0.0373]	−0.0011 [−0.0017, −0.0006]	+0.0251 [+0.0154, +0.0368]
Shared prefix, 1.5B, Arm A	194	+0.1678 [+0.1495, +0.1871]	−0.0043 [−0.0050, −0.0037]	+0.1635 [+0.1450, +0.1826]
Shared prefix, 7B, Arm B	64	+0.0360 [+0.0175, +0.0567]	−0.0009 [−0.0013, −0.0005]	+0.0351 [+0.0168, +0.0550]
Shared prefix, 1.5B, Arm C	200	+0.1814 [+0.1626, +0.2014]	−0.0048 [−0.0055, −0.0041]	+0.1766 [+0.1586, +0.1955]
Reasoning tree (public)	4,880	+0.1668 [+0.1629, +0.1706]	−0.0205 [−0.0213, −0.0196]	+0.1464 [+0.1430, +0.1498]
Self-debugging loop, 7B	164	+0.0092 [−0.0109, +0.0270]	+0.0113 [−0.0032, +0.0289]	+0.0205 [+0.0097, +0.0320]
Self-debugging loop, 1.5B	164	+0.1253 [+0.0967, +0.1533]	−0.0434 [−0.0636, −0.0219]	+0.0819 [+0.0676, +0.0986]
Tree search, 7B	194	+0.0134 [+0.0011, +0.0258]	+0.0079 [+0.0002, +0.0176]	+0.0213 [+0.0118, +0.0315]

F.6 Pooled intervals at the point of decision

A comparison null with a correctly counted outer unit. For i.i.d. question-level differences $D_q \in [-1, 1]$, Hoeffding's inequality (Hoeffding, 1963) gives a fixed-horizon interval $\bar{D} \pm h_Q$ for $\mathbb{E}D_q$, where $h_Q = \sqrt{2\log(2/\delta)/Q}$. At $Q = 10,000$ and $\delta = 0.05$, $h_Q = 0.02716$. The pooled center is $0.05692 > 2h_Q$, so with probability at least 0.95 its interval excludes zero even though the population succ@k contrast is zero. The same interval applied to the run-level differences falsely excludes zero with probability at most 0.05. This follows by applying the inequality about each statistic's own expectation; it does not estimate bootstrap rejection rates. Outer resampling can estimate uncertainty about the pooled center but cannot turn that center into the population succ@k contrast.

Table 38 compares two systems separated by exactly 0.050 in question-averaged succ@4 with eight interval procedures, four of which it prints (all eight: supplementary Table S22), at equal attempts and, in Table S24, at a common total cost with the measured run overhead $\kappa = 2.96$, over the designs listed in each panel. The eight are the attempt-level bootstrap on the pooled attempts; the run-level percentile bootstrap on the run indicators, with and without the $(R-1)/R$ plug-in variance factor; run-level runCS and run-level Clopper–Pearson on the run indicators; a cluster-robust Wald interval, which takes the run as the cluster inside a question and the question as the cluster across questions; and the discordant-pair runCS and Clopper–Pearson intervals of §F.4. With outer question resampling (frame (a)), the tested null of identical laws holds its level: no procedure's type I is far above the

Table 51: Different-procedure comparison null, evaluated by exact enumeration. Every question has $\pi_A = \delta_{1/2}$ and $\pi_B = (15/16)\delta_1 + (1/16)\delta_0$, with $k = n = 4$ and $R = 2$ independent runs per system. Systems and question clusters are independent. Here $\widehat{C}_{\text{run}} = \overline{Y}$ because $n = k$. The two procedures have equal succ@k but different finite-design pooled expectations.

System	succ@4	$\mathbb{E}[\widehat{C}_{\text{pooled}}]$	$\mathbb{E}[\widehat{C}_{\text{run}}]$
A	$\frac{15}{16}$	$\frac{15}{16}$	$\frac{15}{16}$
B	$\frac{15}{16}$	$\frac{891}{896}$	$\frac{15}{16}$
B minus A	0	$\frac{51}{896}$	0

nominal 0.05. Its two systems have identical laws under the null, so their pooled biases cancel. This does not establish validity for every population succ@k comparison. Writing H for a question population, the intended contrast is $\Delta_s = \mathbb{E}_{q\sim H}[s_{Bq} - s_{Aq}]$, whereas pooled scores center on $\Delta_C = \mathbb{E}_{q\sim H}[\mathbb{E}\widehat{C}^{Bq}_{\text{pooled}} - \mathbb{E}\widehat{C}^{Aq}_{\text{pooled}}]$. Table 51 gives different procedures with equal $\Delta_s = 0$ and nonzero Δ_C. Run indicators center on Δ_s under either resampled or fixed questions; valid interval construction is a separate requirement.

When the question set is held fixed — what a leaderboard cell states — a single run's attempt-level interval certifies a difference that is not there in 0.745 ($n = 32$) to 0.931 ($n = 512$) of replicates at $\kappa = 0$, the design effect of cluster sampling (Kerry and Bland, 1998), and the same picture holds on the measured runs of §6.3 (Table 39). The two mechanisms behind that failure price separately, and neither reading rescues the interval. Scored against the exact pass@k^{iid} of fitted laws — the quantity its label names, so that the label is not at issue — the attempt-level interval still covers only 0.846 where 0.95 is nominal, which still combines finite-run centering bias and interval-width error; scored instead against the exact succ@k, the estimand the pooled point estimate sits near, it covers 0.768 (Table 32); on Arm F (Appendix D.3), against a 128-run fresh reference on the 62 headline questions whose within-run correlation is defined, it contains the reference 0.493 of the time against 0.985 for runCS (Table S4), a different design from Arm A's fresh reference of §5.3. Part of the collapse is definitional, in other words, and part is not.

Which interval, when. For a fixed-sample comparison on one question set, the reported cluster-robust Wald construction is near its nominal level, while run-level Clopper–Pearson is often conservative (supplementary Table S22); for a paired comparison, the discordant-pair runCS of §5.2 keeps its validity under optional stopping and targets the sign of the difference; for a claim about the size of a benchmark-level difference under optional stopping, only run-level runCS survives, at a power of essentially zero at eight runs or fewer (Table 38), so anytime validity is paid in runs. Removing the bootstrap's $(R-1)/R$ plug-in variance factor widens its intervals without supplying a finite-sample guarantee.

At a total cost of $N = 512$ attempts, runCS's maximum-power design among the evaluated designs passing the empirical type-I screen (at most 0.05 plus two Monte Carlo standard errors) changes from $(R, n) = (128, 4)$ to $(64, 5)$ when the measured run overhead is charged (supplementary Table S24, panels (c)–(d)). This is a grid comparison, not a global optimum or a calibration guarantee from the screen.

G Extended scope

This section gives the scope statements that §7 summarises, with their numbers and detail.

Theoretical scope. The threshold concerns uniform unbiasedness over the stated class; accuracy and mean-square risk are priced by the attained bias of Theorem A.6, attained by an estimator whose bias depends on the mixing law only through μ_1: $2 \cdot 4^{-k}$ at $R = k - 1$ over classes containing the two-point family on $\{0, 1\}$, $2(w/4)^k$ for a mean known to lie in an interval of width w, and possibly less under a known bound on $\text{Var}(\theta)$, where estimators reading several attempts per run can do better. Condition (T) excludes smooth parametric families such as $\{\text{Beta}(a, b)\}$, whose plug-in stays downward-biased at finite R (Table 52; full grid in Table S10), and Theorem 4.1(b) holds whenever Π contains two mutually absolutely continuous mixing laws with different means, which (T) ensures.

Table 52: Below the threshold the plug-in's bias is nearly an order of magnitude larger than above it, where $\widehat{U}_4$ is unbiased. Empirical bias and RMSE around the uniform-unbiasedness threshold at $k = 4$ on laws fitted to Arm A (194 questions, 2,000 replicates against a known truth), from Table S10.

runs R'	estimator	bias	RMSE
$2 < k$	plug-in $\widehat{P}_4$ (nothing is unbiased)	−0.0789	0.0807
$16 \geq k$	plug-in $\widehat{P}_4$	−0.0090	0.0106
$16 \geq k$	$\widehat{U}_4$	+0.0001	0.0057

The analysis is per question; adversarial adaptivity is covered by the impossibility result, and the latent-model variance accounting of §5.4 assumes Definition 3.1. Cost allocations depend on the fitted laws and on the estimated cost ratio.

The pass@k^{iid} *interval.* Its run-mean version also needs within-run exchangeability. On the tree search the $k = 4$ gap is +0.0134 in completion order and +0.0190 in generation order (Table 26); their difference is not resolved. Containment references are themselves estimates, and the widths of Table 44 price the constructions tried, not a lower bound. Across §6 the first-attempt sequence's containment is lowest, 0.428 (Table 7), on the public reasoning tree, whose exchangeability check fails, so its reference is the noisier first-sibling estimate and containment there mixes reference noise with interval error.

Containment, decision and collapse. The targets of §5.3 are different events, and a noisy held-out estimate moves a narrow interval more than a wide one; two fitted moments do not identify higher-order succ@k, so exact-truth corrections apply only to the fitted laws (Tables 32, 33, 35). The decision experiment calibrates against identical laws in both systems, so it does not test equal succ@k under different dependence (Table 51); coverage collapse is an $n \to \infty$ limsup bound, implying neither equality nor a universal plateau.

Scaffold contrasts. The self-debugging contrast compares the loop as a whole, its own tests included, with fresh attempts; it does not separate the loop's policy from its shared state.

The remaining paragraphs give detail on individual tables and designs.

Arm B's token limit. It is 1024 rather than 512 new tokens because 31.6% of pilot chains truncated at 512, with truncation scored as failure; both truncation rates are reported.

Decision experiment. Systems there are paired by question and run index, with conditionally independent latent draws rather than common random numbers, so induced covariance could change power.

Conditioning and denominators, in full. Arm C and the headline design without the admissibility condition (Table 11) report the unconditional estimand. Undefined correlations occur when all continuations of a question agree: those questions have a measured zero gap and stay in the gap denominator, whereas the dependence quantiles are conditional, and Table 31 reports additional-assumption sensitivity values rather than an unconditional bound. Arm B is restricted to independently pilot-scored questions between 0.05 and 0.95, not the full benchmark. Arms A and B are supporting, nonidentical pipelines beside the headline design. The prefix-length sweep's 50 questions are a fixed mechanical subset and not a random sample; its common-denominator panel has only 7 questions; its rows use their own seeds and prefixes and are unpaired, and Table 21 places five estimands side by side with no fitted curve or monotonicity claim.

Public corpora and scaffolds, in full. Tree labels judge step correctness and usefulness rather than end-to-end task success, and sibling groups are generated only after an incorrect pre-generated step; group size limits that corpus to $k \leq 4$. The corpus carries no batch, timestamp or selection-policy information, so the statement that one problem's sibling groups are independent and identically distributed is a working assumption; the narrowness of that block's intervals is therefore not independent corroboration of the protocol. The observational agent corpus uses the provider's verifier, not re-run here; its cross-run independence remains unresolved from metadata, its single scaffold's questions are selected by a minimum-repetition threshold, and we report the within-corpus variance ratio rather than an aggregate score. Each generated scaffold is one implementation from one model family on

one benchmark, with training exposure unverified, and the tree search is our implementation rather than a published system's artefact: its value prompt is the model judging itself, with no external verifier in pruning. Generated evidence uses open-weight models of at most 9B parameters.