WHEN DOES SHARING A SAMPLED REASONING PREFIX PAY OFF?

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating several reasoning candidates and keeping any correct one raises coverage, the fraction of problems that at least one candidate solves. Continuing one sampled prefix into several candidates makes this cheaper, but every candidate then inherits the prefix's early choices. We ask when sharing a sampled prefix pays off at an equal budget of generated tokens. When continuations keep the decoding settings and generation allowance of independent sampling, sharing the first m tokens saves a fraction $s(m)$ of a second candidate's generation and removes a fraction $\rho(m)$ of its coverage gain, the share of success variance these tokens explain. **On each problem, independent sampling at the same expected budget covers it at least as often as every preset allocation of candidates to sampled prefixes of length m if and only if $\rho(m) \geq s(m)$.** Measuring both fractions from continuations of stored prefixes, under plans fixed before sampling, on grade-school mathematics of intermediate difficulty with four base models and a thinking model, we find that **early tokens resolve success variance faster than they consume generation**: the shortest tested prefixes explain 1.8 to 3.5 times their share of generation. At the same expected budget, **eight candidates sharing one prefix covered significantly fewer problems than independent sampling for every model and branch position tested**. By the time candidates branch, they already share part of their chances of success.

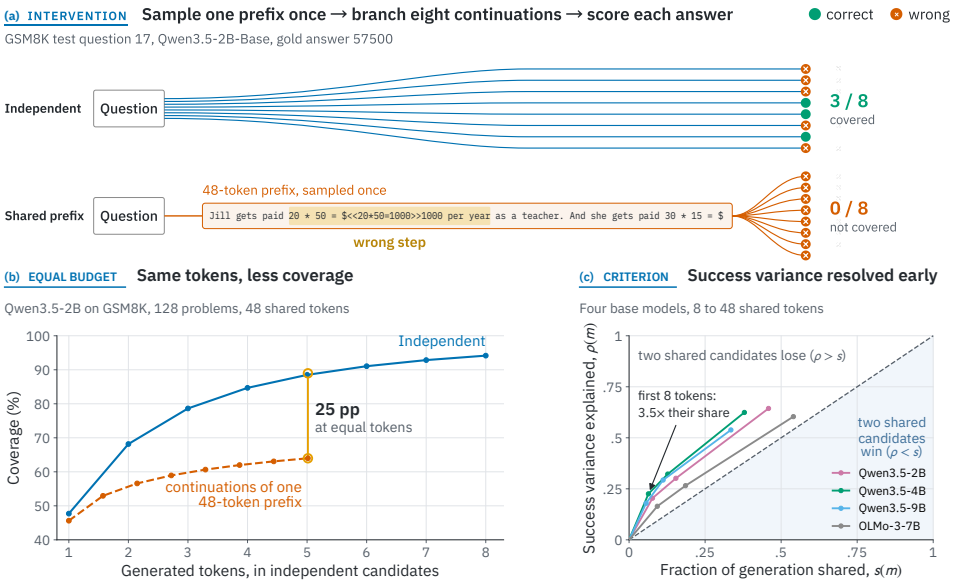

Figure 1: Shared-prefix generation at an equal generated-token budget. (a) Eight independent candidates and eight continuations of one sampled 48-token prefix, whose teaching pay of 20×50 omits the 35 weekly hours (fresh-allowance scan, Appendix E). (b) Coverage against generated tokens, in independent candidates. (c) Share of success variance explained, $\rho(m)$, against share of generation saved, $s(m)$, for four base models; the label gives the largest eight-token ratio.

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.

1 INTRODUCTION

Generating several reasoning candidates for a problem and keeping any correct one is a standard way to solve more problems at inference time (Brown et al., 2024; Snell et al., 2025; Wen et al., 2025; Fu et al., 2026). The set solves a problem whenever one candidate does, so its **coverage**, the fraction of problems solved by at least one of N candidates (pass@N), grows with N (Chen et al., 2021; Singh & Singh, 2026). Because every candidate costs generation, many systems generate a reasoning prefix once and continue it into several candidates. Tree search expands partial solutions (Yao et al., 2023; Ding et al., 2025a; Li et al., 2026b), parallel reasoning forks several continuations from one seed trajectory (Al-Khalili et al., 2026), tree-structured rollouts branch partial trajectories during reinforcement learning (Li et al., 2026a), and serving engines make this reuse cheap by caching the shared tokens (Zheng et al., 2024). Each continuation skips the shared tokens, so the same budget buys more candidates. These methods combine sharing with selection, pruning or training and are evaluated end to end; we isolate the effect of sharing a sampled prefix and relate our analysis to this work in Appendix L.

These candidates are individually as good as independent ones: when continuations use the decoding settings and generation allowance of independent sampling, which we call a matched generation policy, each continuation of a sampled prefix is distributed as an independent sample. What sharing changes is how the candidates depend on one another. Every continuation inherits the choices made in the prefix, such as an interpretation of the problem, an approach, or an intermediate result (Bigelow et al., 2025; Scalena et al., 2026; Li, 2026). In Figure 1(a), the shared 48 tokens contain a wrong step, and all eight continuations reach the same wrong answer, while three of eight independent candidates reach the correct one. Whether sharing pays thus depends on the prefix: tokens that restate the problem save generation at little cost in variety, whereas tokens that choose an approach pass that choice to every candidate. We ask: **at an equal budget of generated tokens, when does sharing a sampled reasoning prefix cover more problems than sampling the candidates independently?**

To answer this question, we describe a prefix by the **probability of eventual success conditioned on that prefix**, the chance that a continuation of it ends in a correct answer (Cheng et al., 2025; Zhou et al., 2026). Under a matched policy, the coverage that sharing loses is determined exactly by how this probability changes along the shared prefix, for any candidate count. With two candidates, the trade reduces to two fractions. Sharing the first m tokens removes a fraction $\rho(m)$ of the second candidate's coverage gain, namely the share of the problem's success variance that the prefix explains, and saves a fraction $s(m)$ of the second candidate's generation. **Sharing pays exactly when $\rho(m) < s(m)$.** The same condition governs every preset allocation of candidates to sampled prefixes of length m: independent sampling at the same expected budget covers a problem at least as often as all of them, for every candidate count, if and only if $\rho(m) \geq s(m)$ on that problem (Theorem 2). **Continuation resampling**, which draws many continuations of stored prefixes, measures $\rho(m)$ (Luo et al., 2024; Ding et al., 2025b). Measuring four base models and a thinking model on grade-school mathematical reasoning, with further models and tasks for breadth, we find:

- **At an equal budget, eight candidates sharing one prefix covered fewer problems than independent candidates.** Across three branch positions for the base models and five for the thinking model, the difference was significant in all 17 comparisons with the same generated tokens, also after Holm's adjustment (Figures 1(b) and 2). Spreading the candidates over two or four prefixes narrows the gap, and in none of the 51 comparisons did shared prefixes cover significantly more; a replay under a hard per-problem budget agrees. In five further settings, including competition problems, the coverage estimates of eight candidates sharing a prefix fell below those of four independent candidates once 8 to 16 tokens were shared, and at 48 tokens every paired interval excludes zero (Section 2).
- **Early tokens resolve success variance faster than they consume generation.** In the base models, the first eight tokens explain 1.8 to 3.5 times as much success variance as their share of generation, and no measured curve of ρ against s falls significantly below the line $\rho = s$ (Figures 1(c) and 4(a)). At 48 tokens, many prefixes without an answer marker already fix the outcome of almost all their continuations. The thinking model, whose success depends largely on concluding within its allowance, shows the same ordering over traces about 52 times longer, and a correct-source proxy for ρ reaches one half within 32 to 49 tokens in ten models from 2B to 35B parameters (Section 4).

- **The same fraction sets the value of continuing a failed prefix.** After one failure, a further continuation of the same prefix succeeds with probability $\mu(1-\rho)$, where μ is the success probability of an independent candidate, so, with fixed costs and decisions that use only the failures, restarting after every failure is optimal exactly when $\rho \geq s$. In the base models, after four failed continuations, a restart succeeded in 40 to 52% of attempts and a further continuation in 4 to 6%. Together, μ and ρ also bound and predict the coverage of larger shared sets (Section 5).
- **A chosen prefix changes what the candidates inherit.** Greedy prefixes raised pass@1 on grade-school problems yet lowered the coverage of the same number of candidates, leaving no successful continuation on some problems that independent sampling solves, whereas short greedy prefixes raised the coverage of one model on competition problems (Section 6).

2 Coverage at an Equal Budget

Sharing a prefix changes both the dependence among candidates and their cost, and we first measure the net effect by comparing candidate sets that cost the same number of generated tokens.

Generation procedures. Under independent sampling, every candidate is generated from the question alone. Under shared-prefix generation with allocation $G \times b$, we sample G prefixes of m generated tokens independently and continue each of them independently b times, giving $N = Gb$ candidates; 1×8 shares one prefix among eight candidates. Continuations use the decoding settings of independent sampling, start from the exact token prefix, and receive the generation allowance that remains after it; a trajectory that has already ended keeps its outcome. Under this **matched generation policy**, every candidate has the distribution of an independent sample (Proposition 4, Appendix B), so sharing leaves pass@1 unchanged and acts on coverage only through the dependence among candidates.

Budget accounting. Both procedures read the same prompt, so we count generated tokens up to the point where a stop string would end generation (Appendix A). Let T be this length for a complete candidate on question q, and let

$$s_q(m) = \frac{\mathbb{E}[\min(T, m) \mid q]}{\mathbb{E}[T \mid q]} \tag{1}$$

be the fraction of a candidate's generation that lies in its first m tokens. An allocation $G \times b$ generates G prefixes and N continuations, so its expected cost equals that of

$$k_{\text{eq}} = G\, s_q(m) + N\big(1 - s_q(m)\big) \tag{2}$$

independent candidates. We compare each allocation with independent sampling of k_{eq} candidates, randomized between the two adjacent integers, which matches expected generated tokens question by question and therefore also over the question set. Because lengths vary across samples, we also replay the recorded token sequences under a hard per-question budget equal to the allocation's expected cost, truncating and grading again the candidate during which it runs out.

Models and data. The base models are Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B and OLMo-3-7B, prompted with four worked GSM8K examples (Cobbe et al., 2021) and sampled at temperature 1.0 and top-p 0.95 with an allowance of 320 tokens, on cohorts of 128 to 192 test questions whose independent samples succeed at rates between 0.15 and 0.85. In these four-prefix measurements, every question receives four sampled prefixes, 32 continuations of each prefix at $m \in \{8, 16, 48\}$, and 32 to 80 independent samples. The thinking model is Qwen3.5-2B in its reasoning mode with an allowance of 8,192 tokens, branched on 128 questions at 2, 5, 10, 20 and 40% of each question's median length over eight screening samples, capped at the allowance, which places its branches at 140 to 2,800 tokens on average. Earlier single-prefix measurements, with one sampled prefix per question at ten positions and Gemma-4-E4B in place of Qwen3.5-4B, complement them (Appendix A).

Estimation. The coverage of a fixed allocation is estimated without bias from its continuations and prefix subsets, and independent coverage at the estimated k_{eq} is a plug-in estimate. Intervals are pointwise 95% paired bootstrap intervals over questions, and the counts of significant comparisons carry no multiplicity adjustment unless stated. The analyses follow a plan fixed before sampling (Appendix C).

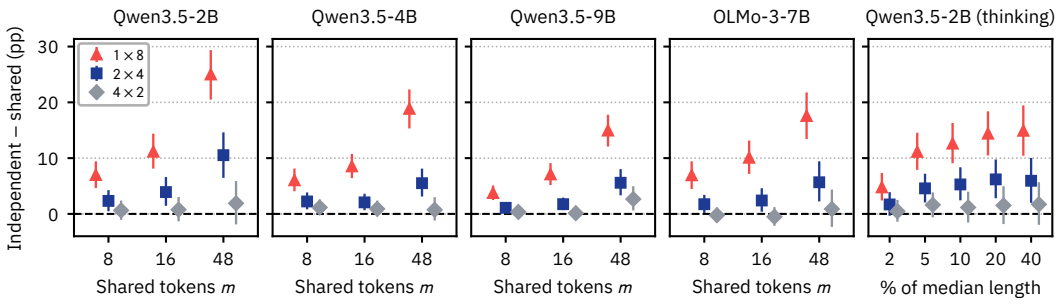

Figure 2: Independent minus shared coverage at the same expected generated-token budget; positive values favor independent sampling. Base models branch at 8, 16 and 48 tokens, the thinking model at fractions of its median length. Bars: 95% paired bootstrap intervals.

Results. Figure 2 compares every allocation with independent sampling at the same expected budget. **Eight candidates sharing one prefix covered significantly fewer problems in all 17 comparisons**, also after Holm's adjustment. For Qwen3.5-2B at 48 tokens, they cover 64.0% of the questions, whereas the same generated tokens buy 5.0 independent candidates on average, which cover 89.0%. Spreading the eight candidates over more prefixes narrows the gap, to significant differences in 16 of 17 comparisons for 2×4 and 1 of 17 for 4×2, because a prefix shared by two candidates both costs and saves less than one shared by eight. No comparison favors sharing significantly, before or after adjustment, and two of the 51 point estimates favor it. The hard-budget replay agrees, with significant differences in 17 of the 17 1×8 comparisons and none favoring sharing (Appendix Tables 4 and 5). A larger mixture-of-experts model, Qwen3.5-35B-A3B, measured in a second pre-registered run, shows the same result at all three positions (Appendix D).

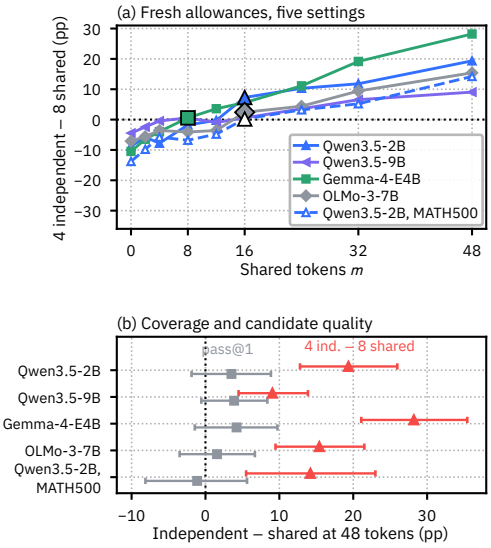

Figure 3: Five settings with fresh allowances: (a) four independent minus eight shared candidates, large markers at the first m whose estimate is above zero; (b) at 48 tokens, the same difference and independent minus shared pass@1, with paired 95% intervals.

The thinking model, which branches after up to 40% of its median length, favors independent sampling significantly in 9 of its 15 comparisons. Shared and independent pass@1, which the matched policy equates, differ by at most 2.1 pp, with all 17 paired intervals containing zero, although every shared estimate is lower (Appendix Table 6); the shortfall does not drive the coverage gaps, since four sampled prefixes continued once each cover 0.6 pp less than four uninterrupted samples on average and 4.1 to 25.6 pp more than one prefix continued four times (Appendix C). The single-prefix measurements agree, with 25 of 40 comparisons favoring independent sampling significantly and none favoring sharing (Appendix C).

Other models and a competition task. An earlier scan, in which every continuation receives a fresh allowance, a generation cap that restarts after the prefix, extends the comparison to Gemma-4-E4B and to competition problems from MATH500 (Hendrycks et al., 2021; Lightman et al., 2024). Under the common-length accounting of Appendix E, eight candidates sharing m tokens cost more than four independent ones whenever $m < 4L/7$, which holds at every measured length because $L \geq 90$. In all five settings, the coverage estimates of eight shared candidates fall below those of four independent ones once 8 to 16 tokens are shared (Figure 3a), and at 48 tokens the four independent candidates lead by 9.1 to 28.2 pp, with every paired interval excluding zero, while pass@1 changes by -4.2 to $+1.1$ pp, with every interval containing zero (Figure 3b; Appendix E).

3 What Sharing Costs and What It Saves

Conditional success. A shared prefix hands every continuation the same chance of success, and we follow that chance along generation. For a question q and generation policy π, let $Z \in \{0, 1\}$ indicate whether a complete candidate is correct, and let $\mu_q = \Pr(Z = 1 \mid q)$. Let $\mathcal{F}_m$ contain the question and the first m generated tokens, or the complete candidate if generation ended earlier. The conditional success probability

$$V_m = \Pr(Z = 1 \mid \mathcal{F}_m), \qquad V_0 = \mu_q, \tag{3}$$

is a martingale in m under the matched policy (Proposition 4 in Appendix B): individual prefixes raise or lower the chance of success, and on average they leave it at μ_q. Given a prefix, N continuations fail independently, each with probability $1 - V_m$, so N candidates that share m tokens cover q with probability

$$C_N(m) = \mathbb{E}\big[1 - (1 - V_m)^N \mid q\big], \qquad C_N(0) = 1 - (1 - \mu_q)^N. \tag{4}$$

The coverage cost of sharing. Let $\phi_N(v) = (1 - v)^N$ be the probability that N continuations of a prefix with success probability v all fail, and let $D_{\phi_N}(b, a) = \phi_N(b) - \phi_N(a) - \phi'_N(a)(b - a) \geq 0$ be its gap above the tangent at a.

Theorem 1 (coverage cost of a shared interval)**.** *Under a matched generation policy, for any integer $N \geq 1$ and positions $m < m'$,*

$$C_N(m) - C_N(m') = \mathbb{E}\big[D_{\phi_N}(V_{m'}, V_m) \mid q\big] \;\geq\; 0. \tag{5}$$

Because V_m is a martingale, the linear term of ϕ_N averages out, so extending a shared prefix costs exactly the expected convexity gap of the change in conditional success probability that all candidates then share. The costs of consecutive intervals add, so the coverage cost of a shared prefix accumulates wherever the conditional success probability moves. For two candidates the gap is the squared change, and starting from the question gives $C_2(0) - C_2(m) = \mathrm{Var}(V_m \mid q)$. The gain of a second independent candidate therefore splits into the part that sharing removes and the part that it keeps (Proposition 5 in Appendix B):

$$\underbrace{\mu_q(1 - \mu_q)}_{\text{gain of a second candidate}} \;=\; \underbrace{\mathrm{Var}(V_m \mid q)}_{\text{removed by sharing}} \;+\; \underbrace{\mathbb{E}[V_m(1 - V_m) \mid q]}_{\text{kept after sharing}}. \tag{6}$$

We write the removed fraction as

$$\rho_q(m) = \frac{\mathrm{Var}(V_m \mid q)}{\mu_q(1 - \mu_q)}, \tag{7}$$

the share of the outcome's variance that the first m tokens explain. It is zero at $m = 0$, never decreases, and reaches one once the prefix determines the outcome. It is also the intraclass correlation between the outcomes of two continuations of one prefix, and in two-stage cluster sampling the variance-optimal number of continuations per prefix exceeds one exactly when $\rho_q < s_q$ (Kish, 1965; Cochran, 1977).

The criterion at an equal budget. A prefix sampled from π saves the fraction $s_q(m)$ of Equation (1) on every continuation after the first, since a prefix and its continuation together are distributed as a complete candidate.

Proposition 1 (two candidates)**.** *Let $0 < \mu_q < 1$. Two candidates sharing m sampled tokens cover q with probability $\mu_q + \big(1 - \rho_q(m)\big)\mu_q(1 - \mu_q)$. Independent sampling with the same expected number of generated tokens, one candidate and a second with probability $1 - s_q(m)$, covers it with probability $\mu_q + \big(1 - s_q(m)\big)\mu_q(1 - \mu_q)$. Sharing covers more if and only if $\rho_q(m) < s_q(m)$.*

A shared second candidate thus keeps the fraction $1 - \rho$ of the value of an independent one at the fraction $1 - s$ of its cost, and independent sampling leads by exactly $(\rho - s)\,\mu(1 - \mu)$. For any additive cost, sharing pays exactly when its expected saving per shared token exceeds $\rho_q(m)/s_q(m)$ times the expected cost per generated token, a condition that per-token costs rising with position make harder to meet and cache reuse across shared continuations (Juravsky et al., 2024) makes easier. When rollout rewards are averaged to estimate μ_q at a fixed budget, as in group baselines, pairs that share a prefix lower the variance only if $\rho_q(m) < s_q(m)/\big(2 - s_q(m)\big)$, a stricter condition (Appendix B).

Theorem 2 (when sharing can pay). *Fix q and m with $0 < \mu_q < 1$. Consider single-level allocations $G \times b$ at position m whose prefixes are sampled independently from π and whose sizes are chosen before any prefix is seen, and compare each with independent sampling of k_{eq} candidates (Equation (2)), randomized between the two adjacent integers independently of the generations, at the same expected number of generated tokens. Independent sampling covers q at least as often as every such allocation if and only if $\rho_q(m) \geq s_q(m)$.*

Necessity follows from Proposition 1. For sufficiency (Appendix B), the failure probabilities $F_b = \mathbb{E}[(1 - V_m)^b \mid q]$ of b continuations of one prefix are log-convex in b, so $F_b \geq (1 - \mu_q)\big(1 - \mu_q + \rho_q\mu_q\big)^{b-1}$, which for $\rho_q \geq s_q$ is at least the failure probability $(1 - \mu_q)\big(1 - (1 - s_q)\mu_q\big)^{b-1}$ of independent candidates at the same expected cost.

From questions to a benchmark. Over a set of questions we summarize both sides by the pooled profiles

$$\rho(m) = \frac{\sum_q \mathrm{Var}(V_m \mid q)}{\sum_q \mu_q(1 - \mu_q)}, \qquad s(m) = \frac{\sum_q s_q(m)\, \mu_q(1 - \mu_q)}{\sum_q \mu_q(1 - \mu_q)}, \tag{8}$$

which weight each question by the gain of its second candidate. With these weights the two-candidate comparison is exact after pooling: averaged over questions, independent sampling leads by $\rho(m) - s(m)$ times the mean of $\mu_q(1-\mu_q)$. Dominance over every allocation remains a statement about each question, which is why Section 2 compares the larger allocations directly.

Scope. Theorem 2 concerns prefixes sampled from the policy and allocations fixed in advance. Both theorems also hold when the branch position is a stopping time τ, such as the first token whose entropy exceeds a threshold, since optional stopping keeps $\mathbb{E}[V_\tau \mid q] = \mu_q$ (Appendix B). A prefix chosen among several by a rule also changes the success probability that its candidates inherit, which Section 6 prices. When $\rho_q < s_q$, the theorem guarantees only that two candidates on one prefix cover more than independent sampling; whether a larger allocation does depends on the whole distribution of V_m, which Section 5 bounds from both sides.

4 Success Variance Along Generation

Section 3 reduces the equal-budget question to two curves that rise from zero to one along the answer: the fraction of success variance that a prefix has resolved, $\rho(m)$, and the fraction of generation it has consumed, $s(m)$.

Estimator. The continuations of a prefix estimate its conditional success probability, and the squared deviation of this estimate from an independent estimate of μ_q, corrected for the noise of finitely many samples, estimates $\mathrm{Var}(V_m \mid q)$ even from one prefix per question. Falling-factorial moments make the numerator and denominator of $\rho(m)$ unbiased, and the lengths of the same samples estimate $s_q(m)$ (Appendix C). The single-prefix measurements, with 64 continuations at each of ten positions, trace each profile densely (Appendix F.4).

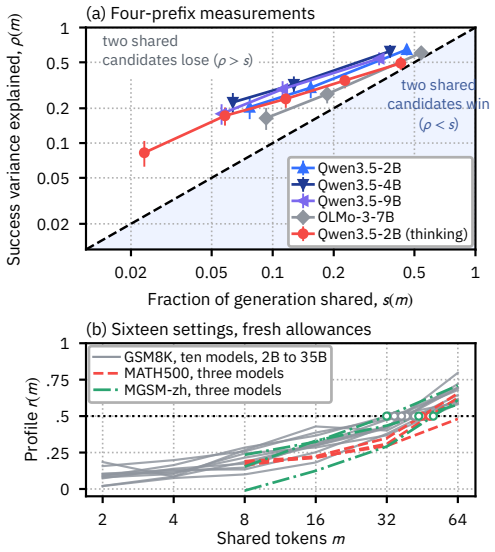

Figure 4: (a) Four-prefix profiles; below $\rho = s$, two shared candidates win on average; bars: 95% intervals. (b) Correct-source profiles $r(m)$ in sixteen settings.

Results. Figures 1(c) and 4(a) plot $\rho(m)$ against $s(m)$ for the four-prefix measurements. Every profile starts above the line $\rho = s$: in the base models, the first eight tokens explain 1.8 to 3.5 times as much success variance as their share of generation, and 11 of the 12 pairs of base model and position lie significantly above the line, none below it, while the interval of the remaining pair, OLMo-3-7B at 48 tokens, includes it. The single-prefix measurements trace profiles of four base models, three of them also measured with four prefixes, through ten

Table 1: Pooled profiles and equal-budget differences at 48 shared tokens (thinking model: 40% of its median length). A positive $\rho - s$, computed before rounding, means that two shared candidates lose on average. Differences are independent minus shared coverage in percentage points at the same expected budget; the last column replays recorded tokens under a hard per-question budget. Bold: 95% interval excludes zero.

					Expected budget			Hard budget
Model	m	$s(m)$	$\rho(m)$	$\rho - s$	1×8	2×4	4×2	1×8
Qwen3.5-2B	48	0.46	0.64	**0.19**	**25.0**	**10.5**	1.9	**24.4**
Qwen3.5-4B	48	0.38	0.62	**0.25**	**18.8**	**5.5**	0.7	**18.0**
Qwen3.5-9B	48	0.33	0.54	**0.20**	**14.9**	**5.6**	**2.7**	**15.3**
OLMo-3-7B	48	0.54	0.60	0.06	**17.5**	**5.7**	0.9	**16.8**
Qwen3.5-2B (thinking)	40%	0.43	0.49	**0.06**	**14.9**	**5.9**	1.7	**14.3**

positions (Appendix Figure 10): the first two tokens explain 2.1 to 6.4 times their share of generation, though at most 1% of prefixes contain the answer marker or have stopped even at eight tokens; the ratio then falls, three profiles stay significantly above the line through 24 tokens and include it from 32 or 96 tokens on, and none falls significantly below it.

For two candidates on one prefix, Section 3 makes the pooled comparison exact, and a check outside the analysis plan bears it out: independent sampling at the same expected budget covers more problems in all 17 settings, significantly in 12, and the gaps match $\rho - s$ times the mean of $\mu_q(1 - \mu_q)$ within 0.8 pp on average (correlation 0.85; Appendix Table 7). For eight candidates the gap depends on more than two moments of V_m (Section 5): OLMo-3-7B, whose interval at 48 tokens includes the line, still loses 17.5 pp there with eight shared candidates (Table 1).

Prefix states. The prefixes themselves show this early resolution (Appendix Figure 11). In the single-prefix measurements at 48 tokens, only 2% to 19% of base-model prefixes contain the explicit answer marker or have stopped, yet among the others, all but at most two of the 64 continuations agree in correctness for 40% to 60%, and none of the 64 continuations succeeds for 8% to 26%, as for the prefix with the wrong step in Figure 1(a). Unparsed answers are about as frequent among continuations as among independent samples in both measurements, and prefixes with no successful continuation fail mostly through wrong answers (Appendix C).

A thinking model. The thinking model answers zero-shot through its chat template over traces of thousands of tokens, so it tests whether early resolution depends on short answers written after worked examples. Without the presence penalty, which the matched policy excludes, half of its traces are still reasoning when the allowance ends, and nearly all traces that conclude are correct (Appendix A.1); its outcome is thus largely whether a trace concludes in time. Its prefixes resolve this outcome faster than they consume generation: the profile lies significantly above the line at every position, and the first 2% of the median length, about 140 tokens, explain 8% of the success variance, 3.5 times their share of generation (Figure 4a).

Other models and tasks. Correct-source profiles extend the measurement to sixteen model–task settings (Figure 4b). The profile $r(m)$ measures how far the continuation success of correct source paths, continued with a fresh allowance, rises above the success of independent samples, normalized so that it equals $\rho(m)$ under a matched policy (Appendix F.5); on the four base models measured both ways, it differs from the matched $\rho(m)$ by 0.05 on average and at most 0.13 through 32 tokens, and by at most 0.15 for Qwen3.5-35B-A3B, with overlapping intervals throughout (Appendix Table 25). On GSM8K, the profiles of ten models from 2B to 35B parameters reach one half within 32 to 49 tokens. On Chinese grade-school problems from MGSM-zh (Shi et al., 2023) they reach it within 32 to 50 tokens, and on MATH500 at 45 tokens or later, with one profile still below one half at 64 tokens. The resolution thus comes later on the competition task.

5 Larger Candidate Sets and Retries

The two moments that settle the two-candidate comparison also bound larger shared sets and decide whether to continue a failed prefix.

Proposition 2 (two-moment bounds). *Fix q and m with $0 < \mu_q < 1$, write $\mu = \mu_q$ and $\rho = \rho_q(m) \in [0, 1]$, and let $c = \mu + \rho(1 - \mu)$. For every integer $b \geq 1$, the probability F_b that b continuations of a sampled prefix all fail satisfies*

$$(1 - \mu)\big(1 - \mu(1 - \rho)\big)^{b-1} \;\leq\; F_b \;\leq\; 1 - \frac{\mu}{c} + \frac{\mu}{c}\,(1 - c)^b, \tag{9}$$

and each bound is attained by a distribution of V_m supported on at most two points: the lower bound by one on $\{\mu(1 - \rho), 1\}$ and the upper bound by one on $\{0, c\}$.

Proposition 3 (retries). *Let $0 < \mu_q < 1$. After $k \geq 1$ failed continuations of a sampled prefix, the next continuation of the same prefix succeeds with a probability h_k that does not increase with k, and $h_1 = \mu_q\big(1 - \rho_q(m)\big)$; a restart succeeds with probability μ_q. If every prefix costs the fraction s of a complete candidate and every continuation the remaining $1 - s$, then among rules that choose between continuing and restarting from the history of failures, restarting after every failure minimizes the expected number of generated tokens until the first success if and only if $\rho_q(m) \geq s$.*

Larger candidate sets. With L_b and U_b the two bounds, an allocation $G \times b$ covers q with probability between $1 - U_b^G$ and $1 - L_b^G$, which settles the comparison whenever independent coverage at the same budget lies outside this range. Between the two ends, a Beta distribution with the same two moments gives a point prediction, using each question's μ_q from its independent samples and the pooled $\rho(m)$ fitted on the other half of the questions. It predicts the coverage of 1×4, 2×4 and 1×8, where two moments alone do not fix it, with a mean absolute error of 2.0 pp and a mean overestimate of 0.7 pp, against 9.9 pp for the prediction that ignores the sharing cost and 2.7 pp for the coverage measured on the other half (Figure 5a; Appendix C.2).

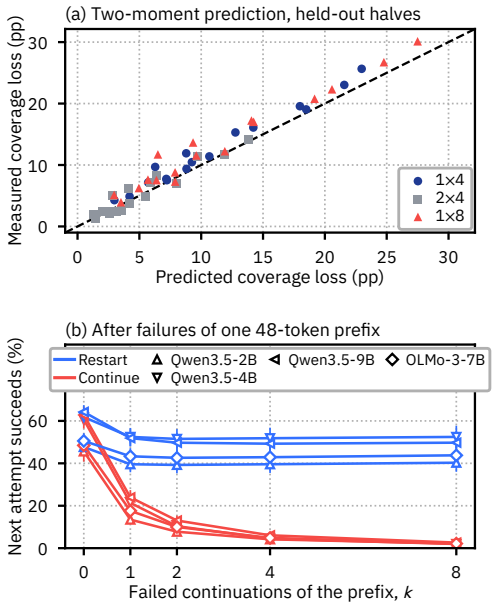

Figure 5: (a) Predicted against measured coverage loss at the same number of candidates, with four or eight continuations per prefix; dashed: identity. (b) Next-attempt success after k failed continuations of one 48-token prefix; bars: 95% intervals.

Continue or restart. Each failure shifts weight toward prefixes with low conditional success, so the success of a further continuation cannot rise with k. Figure 5(b) measures this at 48 tokens, weighting each question by the probability of its failures. After four failures, a further continuation succeeds in 4 to 6% of attempts across base models, whereas a restart succeeds in 40 to 52%. When the four failures come instead from independent candidates of the same questions, continuing a sampled prefix and restarting succeed about equally often (Appendix C.1), so the gap comes from the prefix that the failures share. Within 640 additional generated tokens after four shared failures, restarting after every failure solves 61 to 72 pp more of these questions than continuing the prefix.

6 Selected Prefixes

Search methods choose the prefixes they expand, which changes the success probability that the candidates inherit; Corollary 1 (Appendix I) prices this change.

For a prefix h_q of w tokens with conditional success probability $V_w(h_q)$, chosen without looking at the outcomes of the continuations that evaluate it, N candidates that share it lose

$$C_N(0) - \big[1 - \phi_N(V_w(h_q))\big] = \phi_N'(\mu_q)\,\big(V_w(h_q) - \mu_q\big) + D_{\phi_N}\big(V_w(h_q), \mu_q\big). \tag{10}$$

At fixed N, a deterministic selected prefix improves coverage on question q exactly when $V_w(h_q) > \mu_q$, and because improvements and declines fall on different questions, average pass@1 can rise

while coverage falls. For a randomized rule, Equation (10) compares the gain in success probability, weighted by $-\phi'_N(\mu_q)$, with the expected convexity gap.

Greedy prefixes. The first w tokens of a zero-temperature decode are the simplest selection rule; we continue each greedy prefix 32 times and estimate the coverage of 16 candidates for five models on GSM8K and on a numeric subset of MATH (Figure 6). On GSM8K, greedy prefixes raise pass@1 in all 25 model–length pairs, significantly in 23, yet lower coverage, significantly in 21 of them, by 7.7 to 19.1 pp at 64 tokens (Figure 6b). By Equation (10), a deterministic prefix loses coverage only on questions where it lowers the inherited success probability, and at 64 tokens none of the 32 continuations succeeds on 6.5 to 21.0% of the questions, each of which independent sampling solves at least once in 16 attempts.

On competition problems, short greedy prefixes can raise coverage instead: for Qwen3.5-2B, eight greedy tokens raise the estimated coverage of 16 candidates by 6.3 pp (95% interval 1.3 to 11.1), whereas 64 lower it by 10.9 pp; no other model gains significantly at any length.

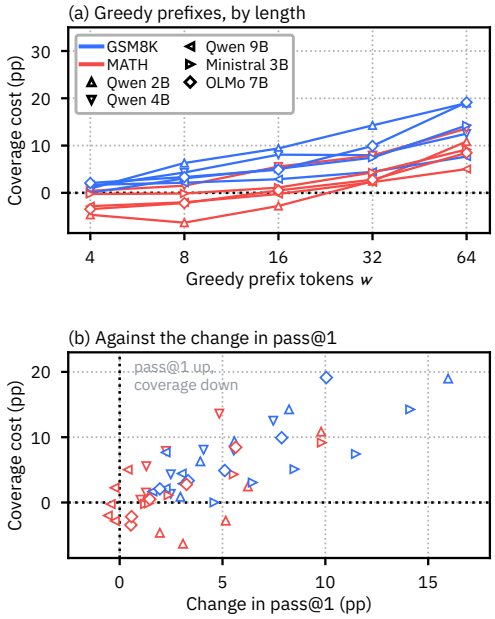

Figure 6: Independent pass@16 minus the coverage of 16 candidates that share the first w greedy tokens: (a) by length; (b) against the change in pass@1, in all fifty settings.

Scored sampled prefixes. Keeping the most confident of two or four sampled prefixes, with every sampled prefix charged, did not offset the sharing cost: for Qwen3.5-35B-A3B, independent sampling at the same budget covered significantly more problems in all 14 settings, and the kept prefix beat a randomly kept one at the same cost significantly in one (Appendix D).

7 Conclusion

In the models we measured, early tokens resolve success variance faster than they consume generation: sharing sampled prefixes brought no significant coverage gain at an equal budget, eight shared candidates covered fewer problems than four independent ones at 48 shared tokens in five further settings, and continuing a failed prefix was far less productive than restarting. The criterion turns these measurements into a procedure: estimate ρ and s from continuations of stored prefixes on pilot questions, which settles the average two-candidate comparison, evaluate larger preset allocations within the available sample counts, and assess a rule that selects prefixes with fresh continuations and a budget that includes the cost of selection.

Our matched-policy measurements cover grade-school mathematics with base models of up to 35B parameters and one thinking model, and the broader measurements reach competition problems and Chinese problems; coverage presumes a verifier that recognizes a correct answer. Under a cost other than generated tokens, the average two-candidate comparison at eight shared tokens favors sharing only if the expected saving per shared token exceeds 1.8 to 3.5 times the expected cost per generated token. Branch positions set by a rule that decides at each token from the tokens generated so far fall under Theorem 2, whereas allocation sizes chosen after seeing a prefix, multi-level trees and continuations forced to differ lie outside it. Majority voting, which needs no verifier (Wang et al., 2023), also favored independent sampling at an equal budget in the one model where we measured it (Appendix D); whether harder tasks move the profile toward the line is the next test of the criterion. **By the time candidates branch, they already share part of their chances of success, and whether sharing a sampled prefix can pay is decided by comparing that share with the share of generation it saves.**

Reproducibility Statement

All experiments use open-weight models at fixed revisions. Appendix A lists prompts, decoding settings, cohorts and budgets, Appendix C the analysis plan fixed before sampling and every deviation from it, and Appendix B the proofs. The supplementary material contains, for both pre-registered runs, the generation scripts, the frozen protocols and analysis plans, the question cohorts, the per-question outcome arrays and the analysis code that regenerates the reported numbers and tables, together with the derived results of the earlier measurements; we will release all generated token sequences.

AI Use Statement

We used large language model assistants to write and check analysis code, to check derivations, and to edit the text. The authors designed the study, verified every derivation and reported number against the generated records, and take responsibility for the content.

References

Zena Al-Khalili, Rafi Hakim, Dietrich Klakow, and Ji-Ung Lee. Fork-think with confidence. *arXiv preprint arXiv:2606.31484*, 2026.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

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

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, 1977.

Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, Xianglong Liu, and Dacheng Tao. Dynamic parallel tree search for efficient LLM reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11233–11252. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.550.

Yuyang Ding, Xinyu Shi, Juntao Li, Xiaobo Liang, Zhaopeng Tu, and Min Zhang. SCAN: Self-denoising Monte Carlo annotation for robust process reward learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations (ICLR)*, 2026.

Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, 1999.

Conor F. Hayes, Elliot Meyerson, Kajetan Schweighofer, Roberto Dailey, Babak Hodjat, Risto Miikkulainen, and Xin Qiu. Beyond the best guess: Improving LLM solution coverage with evolution strategies. *arXiv preprint arXiv:2608.12679*, 2026.

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*, 2021.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. TreeRL: LLM reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12355–12369. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.604.

Sheng Jia, Xiao Wang, and Shiva Prasad Kasiviswanathan. Training large language models to reason in parallel with global forking tokens. In *International Conference on Learning Representations (ICLR)*, 2026.

Ishan Jindal, Sai Prashanth Akuthota, Jayant Taneja, and Sachin Dev Sharma. The path of least resistance: Guiding LLM reasoning trajectories with prefix consensus. In *International Conference on Learning Representations (ICLR)*, 2026.

Jordan Juravsky, Bradley Brown, Ryan Ehrlich, Daniel Y. Fu, Christopher Ré, and Azalia Mirhoseini. Hydragen: High-throughput LLM inference with shared prefixes. *arXiv preprint arXiv:2402.05099*, 2024.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *International Conference on Machine Learning (ICML)*, 2025.

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

Xintong Li, Sha Li, Yuwei Zhang, Changlong Yu, Rongmei Lin, Hongye Jin, Shuyi Guan, Xin Liu, Linwei Li, Qingyu Yin, and Jingbo Shang. Process reward informed tree rollout for effective multi-turn RL. *arXiv preprint arXiv:2607.15610*, 2026a.

Xueyan Li, Johannes Zenn, Ekaterina Fadeeva, Guinan Su, Mrinmaya Sachan, and Jonas Geiping. Efficient test-time inference via deterministic exploration of truncated decoding trees. *arXiv preprint arXiv:2604.20500*, 2026b.

Yingru Li. The variance of thought: Policy variance, critical forks, and local credit assignment. *arXiv preprint arXiv:2608.22467*, 2026.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. TreePO: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Luby, Alistair Sinclair, and David Zuckerman. Optimal speedup of Las Vegas algorithms. *Information Processing Letters*, 47(4):173–180, 1993.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Jiayi Pan, Xiuyu Li, Long Lian, Charlie Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. Learning adaptive parallel reasoning with language models. In *Conference on Language Modeling (COLM)*, 2025.

Gleb Rodionov, Roman Garipov, Alina Shutova, George Yakushev, Erik Schultheis, Vage Egiazarian, Anton Sinitsin, Denis Kuznedelev, and Dan Alistarh. Hogwild! inference: Parallel LLM generation via concurrent attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Daniel Scalena, Sara Candussio, Luca Bortolussi, Elisabetta Fersini, Malvina Nissim, and Gabriele Sarti. Beyond the commitment boundary: Probing epiphenomenal chain-of-thought in large reasoning models. *arXiv preprint arXiv:2606.13603*, 2026.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*, 2023.

Pranav Singh and Prashant Singh. What fixed-rollout pass@k evaluations can identify. *arXiv preprint arXiv:2609.09245*, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Manuel Villén-Altamirano and José Villén-Altamirano. RESTART: A method for accelerating rare event simulations. In *Queueing, Performance and Control in ATM: Proceedings of the 13th International Teletraffic Congress*, pp. 71–76, 1991.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xinglin Wang, Hao Lin, Shaoxiong Feng, Peiwen Yuan, Yiwei Li, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. Share more, search less: Collaborative parallel thinking for efficient test-time scaling. *arXiv preprint arXiv:2605.27030*, 2026.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Hao Wen, Yifan Su, Feifei Zhang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. ParaThinker: Native parallel thinking as a new paradigm to scale LLM test-time compute. *arXiv preprint arXiv:2509.04475*, 2025.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations (ICLR)*, 2025.

Xinyu Yang, Yuwei An, Hongyi Liu, Tianqi Chen, and Beidi Chen. Multiverse: Your language models secretly decide how to parallelize and merge generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (NeurIPS)*, 2023.

Mianjie Yu, Zizhao Mo, Huanyu Qu, Zhirong Qian, Huanle Xu, Cen Li, Zifeng Zhao, Zhi Zhou, Jinhua Zhou, Jun Xie, and Chengzhong Xu. psRL: Efficient training for agentic AI via training-time prefix sharing. *arXiv preprint arXiv:2608.25683*, 2026a.

Youngjun Yu, Sanghwan Jang, and Hwanjo Yu. Difficulty-adaptive tree-structured policy optimization for expanding reasoning coverage in RLVR. *arXiv preprint arXiv:2609.08650*, 2026b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Qiancheng Zhou and Ruizhe Li. Locked at the entrance, open inside: Where RLVR narrows the solution space. *arXiv preprint arXiv:2608.29188*, 2026.

Yuhang Zhou, Yixin Cao, and Guangnan Ye. From correctness to utility: Gain-based prefix evaluation for LLM reasoning. *arXiv preprint arXiv:2606.07190*, 2026.

Jiace Zhu, Yuanzhe Huang, Yingtao Shen, Jie Zhao, and An Zou. Path-consistency with prefix enhancement for efficient inference in LLMs. *arXiv preprint arXiv:2409.01281*, 2024.

A EXPERIMENTAL SETUP

This appendix specifies the models, prompts, decoding settings and question cohorts behind every measurement in the paper: the four-prefix and thinking-model measurements that Sections 2 to 5 analyze, and the earlier measurements that Appendices C and E to K report. Table 2 lists the measurement behind each main-text result.

A.1 FOUR-PREFIX MEASUREMENTS AND THE THINKING MODEL

Base models. Qwen3.5-2B-Base, Qwen3.5-4B-Base, Qwen3.5-9B-Base and OLMo-3-7B-Base (Hugging Face revisions `b1485b2`, `1001bb4`, `68c46c4` and `a81bae4` of `Olmo-3-1025-7B`) receive four worked GSM8K training examples followed by the test question, and generate with temperature 1.0, top-p 0.95, no top-k truncation and a total allowance of 320 new tokens. The answer is the last number after `####` in the text before the next `Question:` marker, and a candidate's useful length counts its tokens up to that marker. A prefix is the exact list of its first m generated token IDs; a continuation receives the remaining $320 - m$ tokens, and a source that ended before m keeps its outcome. The Qwen3.5-2B cohort consists of the 128 questions of the earlier single-prefix measurements; Qwen3.5-4B has 128 new questions, Qwen3.5-9B 64 earlier and 128 new ones, and OLMo-3-7B 64 earlier and 64 new ones. New questions are the next test questions in ascending order whose success rate on eight earlier screening samples lies between 0.15 and 0.85; the screening samples do not enter any estimate. Every question receives four new sampled prefixes, 32 continuations of each at $m \in \{8, 16, 48\}$, and an independent bank of 80 samples (new questions) or 32 samples (earlier questions); the earlier prefixes and banks enter only the comparison with the earlier measurements.

Second pre-registered run. Qwen3.5-35B-A3B-Base (revision `0f08130`), a mixture-of-experts model with 3B active parameters, receives the prompt, decoding settings, allowance, positions and banks of the base models above: four sampled prefixes, 32 continuations of each at $m \in \{8, 16, 48\}$, and an independent bank of 80 samples. It is sampled with vLLM 0.25.1 in bfloat16 over two GPUs instead of Hugging Face Transformers. Its cohort is the first 128 GSM8K test questions in ascending order with two to six correct among eight screening samples; 318 of the 1,319 questions qualify. Its roots also store the log-probability of every sampled token and the top-20 log-probabilities, for the selection analysis of Appendix D.

Thinking model. Qwen3.5-2B (revision `15852e8`) runs in its reasoning mode through its chat template, with the instruction to reason step by step and put the final answer in `\boxed{}`. It generates up to 8,192 tokens with temperature 1.0, top-p 0.95, top-k 20 and no presence or repetition penalty, because such penalties treat prompt and generated tokens differently and would change the policy when a prefix is supplied as a prompt; its model card recommends a presence penalty of 1.5. The answer is the last number in the final `\boxed{}` after the reasoning block; a trace that does not close its reasoning block is incorrect. Every GSM8K test question receives eight screening samples, and the first 128 questions in ascending order with two to six correct samples form the cohort. For question q with median screening length L_q, the positions are $\max(4, \mathrm{round}(fL_q))$ for $f \in \{0.02, 0.05, 0.10, 0.20, 0.40\}$. Each question receives four sampled prefixes, 32 continuations of each at every position, and an independent bank of 80 samples; all token IDs are stored for the hard-budget replay. At this allowance, 50% of the independent samples close the reasoning block, and 96.7% of those are correct; the correct number already appears in the reasoning of 283 of the 284 source traces that reach the allowance without closing it. In screening, 69% of the samples reached the allowance and 502 of the 1,319 questions met the cohort rule; the median screening length of 55 of the 128 cohort questions equals the allowance, so their positions are fractions of 8,192 tokens.

Processing and cutoff. Questions are processed in a random order fixed before sampling, so that a run stopped by the data cutoff keeps the questions claimed first, apart from those still running; only questions with every record complete enter the analysis. Every sample has its own seed, derived from the model, question, prefix and position.

A.2 EARLIER MEASUREMENTS

Table 3 maps the model checkpoints of the earlier measurements to the experiments and their retained question cohorts. The benchmarks are GSM8K (G) (Cobbe et al., 2021), MATH500

Table 2: Measurements behind the main-text results. Planned: the analysis plan was fixed before sampling.

Measurement	Planned	Models and task	Main text	Appendix
Four-prefix, matched policy	yes	Qwen3.5-2B, 4B and 9B, OLMo-3-7B; GSM8K	Figs. 1(b, c), 2, 4(a), 5; Table 1; Sections 2–5	A.1, C
Thinking model, matched policy	yes	Qwen3.5-2B, reasoning mode; GSM8K	Figs. 2, 4(a), 5(a); Table 1; Sections 2, 4, 5	A.1, C
Second run, matched policy	yes	Qwen3.5-35B-A3B; GSM8K	Sections 2, 4, 6	D
Single-prefix, matched policy	no	Qwen3.5-2B and 9B, Gemma-4-E4B, OLMo-3-7B; GSM8K	Sections 2, 4	C, F.3, F.4
Fresh-allowance scan	no	five model–task settings; GSM8K, MATH500	Figs. 1(a), 3; Section 2	E.1
Correct-source profiles	no	sixteen model–task settings; GSM8K, MGSM-zh, MATH500	Fig. 4(b); Section 4	F.5
Greedy prefixes	no	five models; GSM8K, numeric MATH subset	Fig. 6; Section 6	I.2

(M_5) (Hendrycks et al., 2021; Lightman et al., 2024), MGSM-zh (Z) (Shi et al., 2023), and the numeric/fraction-answer MATH subset (M_n) used for the greedy-prefix and temperature experiments. Profile counts refer to questions with a correct source path measured at every position.

Every earlier cohort is prompted with four GSM8K demonstrations. Sampled generation uses top-$p = 0.95$ and temperature 1.0; greedy prefixes use temperature 0, and the temperature interventions use 0.3 or 1.0. Let $B = 320$ tokens on GSM8K and MGSM-zh, and $B = 640$ on the two MATH datasets. Fresh allowances restart after the prefix. The single-prefix measurements use exact token prefixes, top-$k = 0$, the remaining $320 - m$ tokens and absorbing terminal outcomes; their 64-continuation bank measures conditional success and their separate 16-continuation bank measures coverage at $N = 1, 2, 4, 8, 16$, both from the same unselected source of each question. The shared-coverage scan also restarts its allowance; the saved MATH500 records omit the continuation-limit configuration. The scan, prediction, single-prefix and profile cohorts apply the difficulty screen of Section 2, an independent-sample success rate between 0.15 and 0.85, and each keeps the questions complete at every position it uses. Greedy-prefix and temperature experiments use the first 200 questions in each task's fixed order. Appendix F gives the cohort and grid definitions in detail.

The prefix-evaluator comparison uses four cohorts of matched source paths (Appendix H) and broader cohorts of independently sampled candidates (Appendix K.1).

Table 3: Checkpoints, experiments and sampling protocols of the earlier measurements. Profile columns give retained questions; task symbols are defined in the text.

(a) Model checkpoints and experiment coverage

	Profiles (questions)			Coverage / prediction	Single prefix	Greedy / temperature
Model checkpoint	GSM8K	MATH500	MGSM-zh			
Qwen3.5-2B-Base	243	90	107	G, M_5	G	G, M_n
Qwen3.5-4B-Base	283	–	–	–	–	G, M_n
Qwen3.5-9B-Base	278	63	83	G	G	G, M_n
Qwen3.5-35B-A3B-Base	137	–	–	–	–	–
Gemma-4-E2B	268	–	–	–	–	–
Gemma-4-E4B	252	93	–	G	G	–
Gemma-4-12B	287	–	–	–	–	–
Gemma-4-31B	113	–	–	–	–	–
Ministral-3-3B-Base	274	–	–	–	–	G, M_n
OLMo-3-7B-Base	282	–	119	G	G	G, M_n

(b) Sampling and generation allowances

Experiment	Questions	Samples per question or prefix	Token allowance
Profiles	Panel (a)	16 free candidates; 8 continuations per correct source prefix	Fresh B
Shared coverage	100–200	16 free candidates; 8 candidates per shared prefix	Fresh; see text
Coverage prediction	91–193	16 free candidates; 8 continuations per correct and parsed incorrect source prefix	Fresh B
Single prefix	64 or 128	64 continuations for conditional success; 16 for coverage at $N = 1, 2, 4, 8, 16$	Remaining $320 - m$
Greedy prefixes	200	32 continuations per greedy prefix; estimated pass@16	Remaining $B - w$
Temperature	200	16 candidates per sampling condition	Total B across stages

B PROOFS

This appendix proves the statements of Sections 3 and 5; Appendix I proves its corollary where it is stated. Throughout, q is a fixed question, π the generation policy, Y a complete candidate with verifier outcome $Z = v(q, Y) \in \{0, 1\}$, and $(\mathcal{F}_m)_{m \geq 0}$ the filtration generated by q together with the first m generated tokens, where a candidate that terminates before m tokens contributes its complete string; thus $\mathcal{F}_m \subseteq \mathcal{F}_{m'}$ for $m < m'$. Because Z is bounded, every conditional expectation below exists.

Standing assumptions. Complete candidates and the verifier are measurable. Generation stops at a stop string, an end token or a finite allowance, so the length T of a complete candidate is bounded and $0 < L_q = \mathbb{E}[T \mid q] < \infty$. A stopped prefix records that it has stopped. The prefixes of an allocation are independent draws from π, and given the prefixes, all continuations are independent draws from the continuation distribution of π with the remaining allowance and the same stopping rule. Independent candidates, the independent banks and any randomization of candidate counts are independent of these draws. Statements involving $\rho_q(m)$ assume $0 < \mu_q < 1$; for $\mu_q \in \{0, 1\}$ every procedure covers q with probability μ_q.

Proposition 4 (marginal preservation)**.** *Under a matched generation policy, every candidate in shared-prefix generation has distribution $\pi(\cdot \mid q)$. The process $(V_m)_{m \geq 0}$ is a martingale with respect to $(\mathcal{F}_m)_{m \geq 0}$, with $\mathbb{E}[V_m \mid q] = \mu_q$ at every position.*

Proof of Proposition 4. Take the two claims in turn.

Marginal distribution. An autoregressive policy factors any complete sequence $y = (y_{1:m}, y_{m+1:T})$ as

$$\pi(y \mid q) = \pi(y_{1:m} \mid q)\, \pi(y_{m+1:T} \mid q, y_{1:m}).$$

Shared-prefix generation draws $y_{1:m} \sim \pi(\cdot \mid q)$ once and draws each continuation from $\pi(\cdot \mid q, y_{1:m})$ under the decoding settings used for independent generation. The resulting candidate therefore has exactly the density on the right, which is the density of a candidate produced in a single pass. If the prefix terminates before position m, the second factor is a point mass at the empty continuation and the identity still holds. Hence $Y \sim \pi(\cdot \mid q)$ for every candidate and every sharing length, and Z is Bernoulli with parameter $\mu_q = \Pr(Z = 1 \mid q)$.

Martingale property. By (3), $V_m = \mathbb{E}[Z \mid \mathcal{F}_m]$. For $m < m'$ the tower property applied to $\mathcal{F}_m \subseteq \mathcal{F}_{m'}$ gives

$$\mathbb{E}[V_{m'} \mid \mathcal{F}_m] = \mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{F}_{m'}] \mid \mathcal{F}_m\big] = \mathbb{E}[Z \mid \mathcal{F}_m] = V_m,$$

and $0 \le V_m \le 1$ supplies integrability, so $(V_m)_{m \ge 0}$ is a martingale. Since $\mathcal{F}_0$ carries q alone, $V_0 = \mu_q$, and the case $m = 0$ gives $\mathbb{E}[V_{m'} \mid q] = \mu_q$ for every m'. □

Proof of Theorem 1. Given $\mathcal{F}_m$ the N continuations are independent and each fails with probability $1 - V_m$, so all N fail with probability $\phi_N(V_m) = (1 - V_m)^N$ and $C_N(m) = \mathbb{E}[1 - \phi_N(V_m) \mid q]$. For positions $m < m'$,

$$C_N(m) - C_N(m') = \mathbb{E}\big[\phi_N(V_{m'}) - \phi_N(V_m) \mid q\big].$$

By the definition of D_{ϕ_N},

$$\phi_N(V_{m'}) - \phi_N(V_m) = D_{\phi_N}(V_{m'}, V_m) + \phi_N'(V_m)\,(V_{m'} - V_m).$$

The factor $\phi_N'(V_m)$ is $\mathcal{F}_m$-measurable and bounded on $[0, 1]$, so by Proposition 4

$$\mathbb{E}\big[\phi_N'(V_m)(V_{m'} - V_m) \mid \mathcal{F}_m\big] = \phi_N'(V_m)\big(\mathbb{E}[V_{m'} \mid \mathcal{F}_m] - V_m\big) = 0.$$

Taking expectations given q and combining the last three displays gives (5). For $N \ge 2$, $\phi_N''(v) = N(N-1)(1-v)^{N-2} \ge 0$ on $[0, 1]$, so ϕ_N is convex, $D_{\phi_N} \ge 0$, and the loss is nonnegative. At $N = 1$, ϕ_1 is linear and both sides vanish, in agreement with $C_1(m) = \mu_q$. For $N = 2$, $\phi_2'(a) = -2(1 - a)$ and

$$D_{\phi_2}(b, a) = (1 - b)^2 - (1 - a)^2 + 2(1 - a)(b - a) = (b - a)^2,$$

which turns (5) into $C_2(m) - C_2(m') = \mathbb{E}[(V_{m'} - V_m)^2 \mid q]$. For positions $0 = m_0 < m_1 < \cdots < m_K$, applying (5) on each interval and summing gives

$$C_N(0) - C_N(m_K) = \sum_{k=1}^{K} \mathbb{E}\big[D_{\phi_N}(V_{m_k}, V_{m_{k-1}}) \mid q\big],$$

because consecutive intervals share an endpoint; every summand is nonnegative, so the coverage cost accumulates along generation. □

Proposition 5 (two candidates)**.** *For every question and position m, Equation (6) decomposes the gain from a second independent candidate into the coverage lost by sharing and the gain retained after sharing:*

$$C_2(0) - C_2(m) = \mathrm{Var}(V_m \mid q), \qquad C_2(m) - C_1(m) = \mathbb{E}[V_m(1 - V_m) \mid q].$$

Proof of Proposition 5. Conditionally on $\mathcal{F}_m$ the outcome Z is Bernoulli with parameter V_m, so $\mathrm{Var}(Z \mid \mathcal{F}_m) = V_m(1 - V_m)$. Proposition 4 gives $\mathbb{E}[V_m \mid q] = \mu_q$, hence

$$\mathrm{Var}(V_m \mid q) = \mathbb{E}[V_m^2 \mid q] - \mu_q^2, \qquad \mathbb{E}[V_m(1 - V_m) \mid q] = \mu_q - \mathbb{E}[V_m^2 \mid q],$$

and adding the two expressions leaves $\mu_q - \mu_q^2$, which is (6). Equivalently, the two right-hand terms are the explained and unexplained parts in the decomposition of $\mathrm{Var}(Z \mid q)$ over $\mathcal{F}_m$.

The three readings as coverage quantities follow from (4). First, $C_1(m) = \mathbb{E}[V_m \mid q] = \mu_q$ for every m, so $C_2(0) - C_1(0) = (2\mu_q - \mu_q^2) - \mu_q = \mu_q(1-\mu_q)$. Second, $C_2(m) = \mathbb{E}[2V_m - V_m^2 \mid q]$, so

$$C_2(0) - C_2(m) = \mathbb{E}[V_m^2 \mid q] - \mu_q^2 = \mathrm{Var}(V_m \mid q).$$

Third, $C_2(m) - C_1(m) = \mathbb{E}[V_m - V_m^2 \mid q] = \mathbb{E}[V_m(1-V_m) \mid q]$.

The profile $\rho_q(m)$ never decreases because, by the martingale property, $\mathrm{Var}(V_{m'} \mid q) - \mathrm{Var}(V_m \mid q) = \mathbb{E}[(V_{m'} - V_m)^2 \mid q] \geq 0$ for $m < m'$. At the allowance B every candidate has stopped, so $V_B = Z$, $\rho_q(B) = 1$ and $s_q(B) = 1$. □

Lemma 1 (expected cost of an allocation). *For a sampled prefix H of m tokens, let $\ell(H) = \min(T, m)$ be its generated length and $u(H) = \mathbb{E}[T - \ell(H) \mid H, q]$ the expected length of a continuation of H. An allocation $G \times b$ with prefixes $H_1, \ldots, H_G$ generates $\sum_{g=1}^{G} \big(\ell(H_g) + b\,u(H_g)\big)$ tokens in expectation given its prefixes, and*

$$\mathbb{E}[\text{cost of } G \times b \mid q] = L_q\big(G\,s_q(m) + Gb\,(1 - s_q(m))\big) = L_q\,k_{\mathrm{eq}}.$$

Proof. A prefix followed by one continuation is distributed as a complete candidate, so $\mathbb{E}[\ell(H) + u(H) \mid q] = L_q$ and $\mathbb{E}[u(H) \mid q] = L_q - \mathbb{E}[\min(T, m) \mid q] = (1 - s_q(m))L_q$. Summing over the G prefixes and their b continuations each gives the display. No independence between $u(H)$ and the success probability of H is needed. □

Proof of Proposition 1. Write $\mu = \mu_q$, $s = s_q(m)$ and $\rho = \rho_q(m)$. By Lemma 1 with $G = 1$ and $b = 2$, two candidates sharing m tokens cost $(2 - s)L_q$ in expectation, the cost of independent sampling that generates one candidate and a second one with probability $1 - s$. By Proposition 5, the shared pair covers q with probability $C_2(m) = \mu + \mathbb{E}[V_m(1 - V_m) \mid q] = \mu + (1 - \rho)\mu(1 - \mu)$. The independent scheme covers it with probability $\mu + (1 - s)\big(C_2(0) - \mu\big) = \mu + (1 - s)\mu(1 - \mu)$. The difference is $(s - \rho)\,\mu(1 - \mu)$, positive exactly when $\rho < s$. □

Other costs and estimation. The proof uses the cost only through the expected cost of the shared pair. If that pair costs $2 - \sigma$ times the expected cost of a complete candidate for some $\sigma \in [0, 1]$, the same computation with σ in place of s shows that sharing covers more exactly when $\rho < \sigma$. Equivalently, for a candidate of expected cost $\mathbb{E}[C \mid q]$, the expected saving per shared token, $\sigma\,\mathbb{E}[C \mid q]/\mathbb{E}[\min(T, m) \mid q]$, must exceed ρ/s times the expected cost per generated token, $\mathbb{E}[C \mid q]/\mathbb{E}[T \mid q]$. For a per-token cost c_j that does not decrease with the position j, the prefix's share of the expected cost is $\sum_{j \leq m} c_j a_j \big/ \sum_j c_j a_j$ with $a_j = \Pr(T \geq j \mid q)$; it is at most the token share $\sum_{j \leq m} a_j \big/ \sum_j a_j$, because every c_j with $j \leq m$ is at most every c_j with $j > m$. Shared rollouts also estimate μ_q. For two continuations Z_1, Z_2 of one sampled prefix, $\mathrm{Cov}(Z_1, Z_2 \mid q) = \mathbb{E}[V_m^2 \mid q] - \mu_q^2 = \mathrm{Var}(V_m \mid q)$, so their mean is unbiased with variance $\mu(1 - \mu)(1 + \rho)/2$; this is the design effect of two-stage cluster sampling with intraclass correlation ρ (Kish, 1965). The mean of n independent units, each with variance v and expected cost c, has variance $v/n = vc/(nc)$, so at a large budget the design with the smaller product vc estimates μ_q more precisely. This product is $\mu(1 - \mu)(1 + \rho)(2 - s)/2$ for shared pairs, in units of the expected cost of a complete candidate, and $\mu(1 - \mu)$ for independent candidates, so sharing lowers the variance only if $\rho < s/(2 - s)$; with b continuations per prefix, the shared product divided by the independent one is $\big(1 + (b - 1)\rho\big)\big(s + b(1 - s)\big)/b$. Over real b this ratio is convex and smallest at $b^* = \sqrt{s(1 - \rho)/\big((1 - s)\rho\big)}$, the optimum of two-stage sampling (Cochran, 1977), which exceeds one exactly when $\rho < s$; two continuations per prefix beat one exactly when $b^* > \sqrt{2}$, which is the condition $\rho < s/(2 - s)$. Since $s/(2 - s) \leq s$, whenever sharing does not pay for coverage it does not pay for estimation either.

Branch positions chosen from the prefix. Let τ be a stopping time of $(\mathcal{F}_m)$, bounded by the allowance: whether $\tau = m$ is decided by the first m tokens, as for the first token whose entropy exceeds a threshold. Optional stopping gives $\mathbb{E}[V_\tau \mid \mathcal{F}_\sigma] = V_\sigma$ for stopping times $\sigma \leq \tau$ and $\mathbb{E}[V_\tau \mid q] = \mu_q$, and a prefix cut at τ followed by a continuation is a draw from π, because $\{\tau = m\} \in \mathcal{F}_m$ and the factorization in the proof of Proposition 4 applies on each event $\{\tau = m\}$.

The proofs of Theorems 1 and 2 and of Lemma 1 therefore go through with m replaced by τ, $\rho_q(\tau) = \mathrm{Var}(V_\tau \mid q)/\big(\mu_q(1-\mu_q)\big)$ and $s_q(\tau) = \mathbb{E}[\min(T,\tau) \mid q]/\mathbb{E}[T \mid q]$. Rules that look beyond the prefix, such as continuing several candidates before choosing where to branch, do not define stopping times.

Proof of Theorem 2. Necessity. If $\rho_q(m) < s_q(m)$, the allocation 1×2 covers q with higher probability than independent sampling at the same expected cost, by Proposition 1.

Sufficiency. Write $\mu = \mu_q$, $a = 1-\mu$, $s = s_q(m)$, $\rho = \rho_q(m)$, and $F_b = \mathbb{E}[(1-V_m)^b \mid q]$ for integers $b \geq 0$. Since $\mu < 1$, $\Pr(V_m < 1) > 0$ and $F_b > 0$.

Step 1: log-convexity. By the Cauchy–Schwarz inequality, $F_b^2 = \mathbb{E}\big[(1-V_m)^{(b-1)/2}(1-V_m)^{(b+1)/2}\big]^2 \leq F_{b-1}F_{b+1}$, so the ratios F_{b+1}/F_b do not decrease in b and $F_b = F_1 \prod_{j=1}^{b-1} F_{j+1}/F_j \geq F_1(F_2/F_1)^{b-1}$.

Step 2: the second moment. $F_1 = a$ and $F_2 = a^2 + \mathrm{Var}(V_m \mid q) = a^2 + \rho\mu a$, so $F_2/F_1 = a + \rho\mu \geq a + s\mu = 1-(1-s)\mu$ whenever $\rho \geq s$. Hence $F_b \geq a\big(1-(1-s)\mu\big)^{b-1}$.

Step 3: one prefix. A prefix with b continuations costs $\big(1+(b-1)(1-s)\big)L_q$ in expectation (Lemma 1). Independent sampling that generates one candidate and $b-1$ further candidates, each independently with probability $1-s$, has the same expected cost and fails with probability $a\big(s+(1-s)a\big)^{b-1} = a\big(1-(1-s)\mu\big)^{b-1} \leq F_b$.

Step 4: several prefixes. The G prefixes of an allocation are sampled independently and their continuations are conditionally independent, so the allocation fails with probability F_b^G. Running G independent copies of the scheme of Step 3 costs the same in expectation and fails with probability at most F_b^G. Its number of candidates is $K = G + \mathrm{Binomial}\big(G(b-1), 1-s\big)$, with $\mathbb{E}[K] = k_{\mathrm{eq}}$ of Equation (2).

Step 5: integer candidate counts. For $k \geq 0$ let $n = \lfloor k \rfloor$ and $\theta = k - n$; independent sampling of k candidates randomized between adjacent integers generates $K_* = n + \mathrm{Bernoulli}(\theta)$ candidates, with this randomization independent of all generations, and fails with probability $\mathbb{E}[a^{K_*}] = (1-\theta)a^n + \theta a^{n+1}$. Because $j \mapsto a^j$ is convex, $a^j \geq a^n + (j-n)(a^{n+1}-a^n)$ for every integer $j \geq 0$. Hence every integer-valued count K, independent of the generations, with $\mathbb{E}[K] = k$ satisfies $\mathbb{E}[a^K] \geq (1-\theta)a^n + \theta a^{n+1} = \mathbb{E}[a^{K_*}]$. Applied to the count of Step 4 with $k = k_{\mathrm{eq}}$, independent sampling of K_* candidates fails with probability at most F_b^G at the same expected cost. The argument holds for all integers $G, b \geq 1$, and for predetermined groups of unequal sizes it applies to each group before multiplying. □

Proof of Proposition 3. Keep the notation of the previous proof, set $F_0 = 1$, and let $Z_1, Z_2, \ldots$ be the outcomes of continuations of one sampled prefix. Given $\mathcal{F}_m$ they are independent Bernoulli variables with parameter V_m; the probabilities below condition on q and on the observed failures and average over the sampled prefix, so that

$$h_k = \Pr(Z_{k+1} = 1 \mid Z_1 = \cdots = Z_k = 0) = \frac{\mathbb{E}[V_m(1-V_m)^k \mid q]}{F_k} = 1 - \frac{F_{k+1}}{F_k}, \qquad h_0 = \mu,$$

which does not increase in k by Step 1 above. For $k = 1$, $\mathbb{E}[V_m(1-V_m) \mid q] = \mu(1-\mu)(1-\rho)$ by Equation (6), so $h_1 = \mu(1-\rho)$. A restart draws a new candidate independently of the failed prefix and succeeds with probability μ.

For the cost statement, measure cost in complete candidates, so that a prefix and its first continuation cost one and every further continuation costs $1-s$. Consider any rule that decides after each failure, from the history of failures and independent randomization, whether to continue the current prefix or to start a new one, and that stops at the first success. Suppose $\rho \geq s$. A new attempt costs one and succeeds with probability μ whatever the history. A further continuation after $j \geq 1$ failures of its prefix costs $1-s$ and succeeds with conditional probability $h_j \leq h_1 = \mu(1-\rho) \leq \mu(1-s)$. In both cases the success probability of an attempt is at most μ times its cost. Summing over attempts, which is justified by Tonelli's theorem, gives $1 = \Pr(\text{eventual success}) \leq \mu\,\mathbb{E}[\text{total cost}]$, so every rule costs at least $1/\mu$ in expectation, which restarting after every failure attains. Suppose instead $\rho < s$.

The rule that continues each prefix once after its first failure and then restarts repeats independent cycles with expected cost $D = 1+(1-s)(1-\mu)$ and success probability $C = \mu+\mu(1-\mu)(1-\rho)$, so its expected cost is D/C by the renewal equation $J = D + (1-C)J$, and $D/C < 1/\mu$ exactly when $\rho < s$. Restarting after every failure is therefore optimal if and only if $\rho \geq s$. The argument needs fixed costs: with variable lengths, the cost of continuing depends on the prefix through $u(H)$ of Lemma 1, which may vary with its success probability. □

Proof of Proposition 2. Let $f(v) = (1-v)^b$ on $[0,1]$, so that $F_b = \mathbb{E}[f(V_m) \mid q]$, and recall $\mathbb{E}[V_m \mid q] = \mu$ and $\mathbb{E}[V_m^2 \mid q] = \mu^2 + \rho\mu(1-\mu)$.

Lower bound. Step 1 of the proof of Theorem 2 gives $F_b \geq F_1(F_2/F_1)^{b-1} = (1-\mu)\big(1-\mu(1-\rho)\big)^{b-1}$. If V_m equals $\mu(1-\rho)$ with probability $(1-\mu)/(1-\mu(1-\rho))$ and one otherwise, it has the required two moments and $F_{j+1}/F_j = 1-\mu(1-\rho)$ for every $j \geq 1$, so the bound is attained.

Upper bound. For $b \leq 2$ both sides are equal, because $F_1 = 1-\mu$ and $F_2 = (1-\mu)^2 + \rho\mu(1-\mu)$ are fixed by the two moments. For $b \geq 3$, $f'''(v) = -b(b-1)(b-2)(1-v)^{b-3} \leq 0$ on $[0,1]$. Let g be the quadratic polynomial with $g(0) = f(0)$, $g(c) = f(c)$ and $g'(c) = f'(c)$, where $c = \mu + \rho(1-\mu) \in [\mu, 1]$. The Hermite interpolation remainder gives, for every $v \in [0,1]$, some $\xi_v \in (0,1)$ with

$$f(v) - g(v) = \tfrac{1}{6} f'''(\xi_v)\, v\,(v-c)^2 \;\leq\; 0.$$

Hence $F_b \leq \mathbb{E}[g(V_m) \mid q]$. A quadratic has an expectation fixed by the first two moments, and the distribution with $\Pr(V_m = c) = \mu/c$ and $\Pr(V_m = 0) = 1-\mu/c$ has mean μ and second moment $c\mu = \mu^2 + \rho\mu(1-\mu)$, so $\mathbb{E}[g(V_m) \mid q]$ equals the expectation of g under it. On its support $g = f$, which gives $F_b \leq 1 - \mu/c + (\mu/c)(1-c)^b$, attained by this distribution.

Allocations and pooling. The G prefixes of an allocation are independent, so the allocation fails with probability F_b^G, and its coverage lies between one minus the Gth powers of the two bounds. Step 1 of the proof of Theorem 2 holds for any distribution of V_m on $[0,1]$, including the mixture over questions and sampled prefixes; with the pooled probabilities F_1 and F_2 that one and two continuations of a sampled prefix fail, pooled coverage of N candidates sharing one prefix is at most $1 - F_1(F_2/F_1)^{N-1}$; for several prefixes, benchmark failure is the average of $F_b(q)^G$ over questions, which the pooled moments do not determine. □

B.1 Coverage gains relative to one candidate

Let $\overline{C}_N(m) = \mathbb{E}_q[C_N(m)]$ and assume that $\overline{C}_N(0) - \overline{C}_1(0) > 0$. Under marginal preservation, the fraction of the independent multi-candidate gain removed by sharing is

$$R_N(m) = \frac{\overline{C}_N(0) - \overline{C}_N(m)}{\overline{C}_N(0) - \overline{C}_1(0)} = 1 - \frac{\overline{C}_N(m) - \overline{C}_1(m)}{\overline{C}_N(0) - \overline{C}_1(0)}. \tag{11}$$

The equality uses $\overline{C}_1(m) = \overline{C}_1(0)$. For $N = 2$, Equation (6) makes $R_2(m)$ equal to the pooled profile $\rho(m)$.

Table 17(a) estimates the second expression from each arm's measured single-candidate success rate:

$$\widehat{R}_N(m) = 1 - \frac{\widehat{C}_N(m) - \widehat{C}_1(m)}{\widehat{C}_N(0) - \widehat{C}_1(0)}.$$

It quantifies the reduction in the additional coverage obtained from $N-1$ more candidates. Because the empirical single-candidate rates can differ, this estimate also accounts for their measured change, whereas normalizing only the total coverage difference uses the first expression.

C Equal-Budget Comparisons and Retries in Detail

Analysis plan. The four-prefix and thinking-model measurements follow an analysis plan fixed before sampling. It specifies the estimators below, the handling of incomplete questions (only questions with all records enter; questions are processed in a random order fixed in advance, so that a run interrupted by the cutoff keeps the questions claimed first, apart from those still running,

Table 4: All pre-registered equal-budget comparisons at the same expected generated tokens: independent minus shared coverage in percentage points, with pointwise 95% paired bootstrap intervals, and the pooled profile values at each position. The comparisons share questions and prefixes, and no multiplicity adjustment is applied.

Model	m	$s(m)$	$\rho(m)$	ρ/s	1×8	2×4	4×2
Qwen3.5-2B	8	0.08	0.20	2.7	7.0 [4.6, 9.4]	2.3 [0.5, 4.3]	0.6 [−1.1, 2.4]
Qwen3.5-2B	16	0.15	0.30	2.0	11.1 [8.1, 14.4]	3.9 [1.5, 6.6]	0.8 [−1.3, 3.0]
Qwen3.5-2B	48	0.46	0.64	1.4	25.0 [20.5, 29.4]	10.5 [6.5, 14.6]	1.9 [−1.9, 5.9]
Qwen3.5-4B	8	0.06	0.23	3.5	6.0 [4.1, 8.1]	2.2 [0.8, 3.8]	1.2 [−0.1, 2.6]
Qwen3.5-4B	16	0.13	0.32	2.5	8.5 [6.4, 10.8]	2.1 [0.7, 3.6]	0.9 [−0.4, 2.4]
Qwen3.5-4B	48	0.38	0.62	1.6	18.8 [15.3, 22.3]	5.5 [3.1, 8.1]	0.7 [−1.2, 3.0]
Qwen3.5-9B	8	0.06	0.18	3.2	3.7 [2.5, 5.1]	1.1 [0.4, 1.8]	0.3 [−0.3, 1.0]
Qwen3.5-9B	16	0.11	0.29	2.6	7.1 [5.2, 9.1]	1.7 [0.7, 2.9]	0.1 [−0.7, 0.9]
Qwen3.5-9B	48	0.33	0.54	1.6	14.9 [12.1, 17.8]	5.6 [3.3, 8.0]	2.7 [0.6, 5.0]
OLMo-3-7B	8	0.09	0.16	1.8	6.9 [4.5, 9.4]	1.7 [0.2, 3.4]	−0.2 [−1.5, 1.1]
OLMo-3-7B	16	0.19	0.27	1.4	10.1 [7.2, 13.1]	2.4 [0.4, 4.6]	−0.5 [−2.1, 1.2]
OLMo-3-7B	48	0.54	0.60	1.1	17.5 [13.4, 21.8]	5.7 [2.2, 9.4]	0.9 [−2.3, 4.4]
Qwen3.5-2B (thinking)	2%	0.02	0.08	3.5	4.8 [2.4, 7.3]	1.7 [−0.3, 3.9]	0.5 [−1.4, 2.5]
Qwen3.5-2B (thinking)	5%	0.06	0.17	3.0	11.1 [7.9, 14.5]	4.6 [2.1, 7.2]	1.6 [−0.6, 3.9]
Qwen3.5-2B (thinking)	10%	0.12	0.24	2.1	12.6 [9.1, 16.3]	5.3 [2.4, 8.4]	1.2 [−1.5, 4.0]
Qwen3.5-2B (thinking)	20%	0.23	0.35	1.5	14.4 [10.5, 18.4]	6.2 [2.8, 9.8]	1.5 [−1.8, 5.0]
Qwen3.5-2B (thinking)	40%	0.43	0.49	1.1	14.9 [10.4, 19.5]	5.9 [2.0, 10.0]	1.7 [−2.0, 5.7]

which can favor shorter generations), and the rule that every pre-registered comparison is reported in either direction. The analysis departs from the plan in four respects: the prefix share $s_q(m)$ averages $\min(T, m)$ over the four sampled prefixes of the question, which is also the quantity entering the cost; the hard-budget replay counts useful tokens in both procedures; the retry analysis uses $k \in \{0, 1, 2, 4, 8\}$ instead of $\{1, 4\}$; and the thinking model's retry budget is four independent candidates' worth of tokens instead of 640 tokens. Every planned question of every cohort was completed before the data cutoff.

Estimators. For a prefix whose K continuations contain F failures, and the falling factorial $(x)_b = x(x-1)\cdots(x-b+1)$,

$$\widehat{(1-v)^b} = \frac{(F)_b}{(K)_b}, \qquad b \le K, \tag{12}$$

estimates the probability that b continuations all fail without bias. Allocations $G \times b$ average the product of these estimates over all subsets of G of the question's four prefixes, which is unbiased because different prefixes are independent. Independent coverage at k candidates uses the same estimator on the bank of independent samples, interpolated linearly between the two integers adjacent to k_{eq}. The denominator $\mathbb{E}[T \mid q]$ of $s_q(m)$ is the mean useful length of the independent bank. Coverage estimates at a fixed candidate count are unbiased; independent coverage at the estimated k_{eq} is a plug-in estimate whose finite-sample bias has no established sign. The profile $\rho(m)$ combines per-prefix unbiased estimates of $(V_m - \mu_q)^2$ from the continuation bank and the independent bank. All intervals are pointwise 95% percentile intervals from 4,000 paired bootstrap resamples of questions, which redraw a question together with all of its prefixes, continuations and independent samples.

All comparisons. Tables 4 and 5 list the 51 pre-registered comparisons with their intervals under the two accountings. Under the hard-budget replay, each question receives a budget equal to the allocation's expected cost, and 64 fixed random orders per question determine which prefixes and samples are used. The shared arm pays for each of its G prefixes once, before any continuation, and then draws continuations in turn across the prefixes; the independent arm draws complete samples. Both arms stop when the budget runs out, and the candidate during which it runs out is graded on the tokens generated up to that point.

Robustness. Two checks outside the analysis plan concern the estimated budget and the number of comparisons. The length of independent samples, which sets k_{eq}, and their coverage come from the same bank; estimating the length from one half of the bank and the coverage from the other, averaged over both assignments, moves the 51 differences by at most 0.6 pp over 21 random splits,

Table 5: The comparisons of Table 4 under the hard-budget replay of recorded tokens: independent minus shared coverage in percentage points, with pointwise 95% paired bootstrap intervals.

Model	m	1×8	2×4	4×2
Qwen3.5-2B	8	6.9 [4.5, 9.3]	2.6 [0.7, 4.7]	0.3 [−1.4, 2.1]
Qwen3.5-2B	16	11.0 [7.8, 14.3]	4.1 [1.6, 6.7]	0.4 [−1.8, 2.6]
Qwen3.5-2B	48	24.4 [19.7, 29.0]	11.3 [7.2, 15.5]	1.2 [−2.4, 5.0]
Qwen3.5-4B	8	5.9 [3.9, 8.2]	2.5 [0.9, 4.4]	1.0 [−0.3, 2.6]
Qwen3.5-4B	16	8.2 [6.0, 10.5]	1.8 [0.4, 3.4]	0.8 [−0.5, 2.3]
Qwen3.5-4B	48	18.0 [14.4, 21.5]	5.0 [2.4, 7.7]	0.3 [−1.8, 2.8]
Qwen3.5-9B	8	3.9 [2.7, 5.3]	1.2 [0.4, 2.1]	0.4 [−0.4, 1.1]
Qwen3.5-9B	16	7.3 [5.4, 9.4]	1.8 [0.6, 3.1]	0.3 [−0.6, 1.1]
Qwen3.5-9B	48	15.3 [12.4, 18.3]	5.5 [3.2, 8.1]	2.6 [0.5, 4.9]
OLMo-3-7B	8	7.5 [5.0, 10.1]	1.8 [0.1, 3.5]	−0.2 [−1.6, 1.3]
OLMo-3-7B	16	10.5 [7.4, 13.8]	2.3 [0.1, 4.7]	−0.5 [−2.4, 1.4]
OLMo-3-7B	48	16.8 [12.3, 21.1]	4.0 [0.3, 7.8]	−0.1 [−3.4, 3.4]
Qwen3.5-2B (thinking)	2%	5.3 [2.8, 8.0]	1.8 [−0.3, 4.1]	0.7 [−1.3, 2.8]
Qwen3.5-2B (thinking)	5%	11.0 [7.7, 14.6]	4.4 [1.9, 7.2]	1.1 [−1.2, 3.5]
Qwen3.5-2B (thinking)	10%	12.7 [9.2, 16.5]	4.6 [1.7, 7.8]	0.9 [−1.9, 3.8]
Qwen3.5-2B (thinking)	20%	14.6 [10.7, 18.7]	5.5 [2.1, 9.1]	0.5 [−3.0, 4.1]
Qwen3.5-2B (thinking)	40%	14.3 [9.8, 18.8]	4.3 [0.2, 8.5]	0.0 [−3.9, 4.3]

changes the significance of at most two of them, all borderline, and makes none favor sharing. Holm's step-down procedure at family-wise level 0.05, applied to bootstrap p-values $2\min\{\Pr^*(d \leq 0), \Pr^*(d \geq 0)\}$ from the same 4,000 resamples, keeps 25 of the 51 comparisons and 17 of the 17 1×8 comparisons.

Profiles and the matched-policy check. Table 6 gives, for every position of the four-prefix measurements, the interval of $\rho(m) - s(m)$ that places the profile relative to the line in Section 4, and the difference between the $\mathrm{pass@1}$ of shared continuations and of independent samples. Every $\mathrm{pass@1}$ interval contains zero, but all 17 point estimates are negative. Pooled over the four base models, a check outside the analysis plan, the difference is $-1.7\,\mathrm{pp}$ at 48 tokens (95% interval -3.3 to -0.1), and its intervals at 8 and 16 tokens contain zero; the shortfall also appears for Qwen3.5-35B-A3B, generated with vLLM (Appendix D), and its cause has not been identified. The shortfall does not produce the coverage gaps. Four sampled prefixes continued once each are independent candidates that pass through the same interruption and resumption as shared continuations; the coverage of four uninterrupted samples minus theirs ranges from -0.5 to $2.3\,\mathrm{pp}$, $0.6\,\mathrm{pp}$ on average, with an interval excluding zero in 1 of 17 settings, whereas one prefix continued four times, through the same procedure, covers 4.1 to $25.6\,\mathrm{pp}$ less than the four resumed independent candidates, significantly in all 17 settings. In the base models, resumed candidates also match uninterrupted ones in mean useful length, within 1.9 tokens at every position. Scaling every μ_q by a shift of the measured size lowers the Beta-model coverage of eight shared candidates (Appendix C.2) by at most $1.8\,\mathrm{pp}$, against eight-candidate gaps of 14.9 to $25.0\,\mathrm{pp}$ at 48 tokens. In the earlier single-prefix measurements, one of the four $\mathrm{pass@1}$ intervals at 48 tokens excludes zero, for OLMo-3-7B (Appendix F.3).

Two candidates. Pooled over questions, independent sampling leads one prefix with two continuations by $\rho(m) - s(m)$ times the mean of $\mu_q(1 - \mu_q)$ at the same expected cost (Section 3). Table 7, outside the analysis plan, compares this value, computed from the estimates of $\rho(m)$ and $s(m)$, with the difference measured by the estimators of the pre-registered comparisons. Across the 17 settings the two differ by $0.8\,\mathrm{pp}$ on average, with correlation 0.85, and all 17 paired intervals of their difference contain zero. Because the two sides agree in expectation, the comparison checks the estimators and the cost accounting on the records. The measured gap exceeds the predicted one in 15 of the 17 settings and by $0.8\,\mathrm{pp}$ on average over all of them, which is consistent with the $\mathrm{pass@1}$ shortfall of the shared continuations above: it lowers the measured shared coverage and does not enter the prediction.

Answer formats. Table 8 counts unparsed answers in the four-prefix measurements. Independent samples and continuations fail to parse at similar rates, 1.3% to 2.4% for the Qwen models and 11% to 12% for OLMo-3-7B, whose answers often end with a sentence such as "The answer is 100." without the #### marker. Prefixes whose continuations are mostly unparsed are rare, at most 3.5%

Table 6: Four-prefix profiles relative to the line $\rho = s$, and the matched-policy check: $\rho(m) - s(m)$, computed before rounding, and shared minus independent pass@1 in percentage points, each with a pointwise 95% paired bootstrap interval; n is the number of questions.

Model	m	n	$s(m)$	$\rho(m)$	$\rho(m) - s(m)$	Shared − independent pass@1
Qwen3.5-2B	8	128	0.08	0.20	0.127 [0.085, 0.175]	−0.29 [−2.67, 2.11]
Qwen3.5-2B	16	128	0.15	0.30	0.148 [0.092, 0.206]	−0.49 [−3.39, 2.31]
Qwen3.5-2B	48	128	0.46	0.64	0.185 [0.118, 0.256]	−2.09 [−5.57, 1.34]
Qwen3.5-4B	8	128	0.06	0.23	0.162 [0.120, 0.207]	−1.35 [−3.54, 0.79]
Qwen3.5-4B	16	128	0.13	0.32	0.195 [0.143, 0.248]	−0.88 [−3.32, 1.62]
Qwen3.5-4B	48	128	0.38	0.62	0.245 [0.173, 0.319]	−1.16 [−4.22, 1.99]
Qwen3.5-9B	8	192	0.06	0.18	0.123 [0.088, 0.160]	−1.00 [−2.72, 0.69]
Qwen3.5-9B	16	192	0.11	0.29	0.181 [0.134, 0.231]	−0.99 [−2.86, 0.94]
Qwen3.5-9B	48	192	0.33	0.54	0.204 [0.138, 0.271]	−1.46 [−3.93, 1.14]
OLMo-3-7B	8	128	0.09	0.16	0.072 [0.038, 0.108]	−0.02 [−1.87, 1.88]
OLMo-3-7B	16	128	0.19	0.27	0.080 [0.035, 0.129]	−0.79 [−3.18, 1.60]
OLMo-3-7B	48	128	0.54	0.60	0.063 [−0.004, 0.138]	−2.09 [−5.49, 1.46]
Qwen3.5-2B (thinking)	2%	128	0.02	0.08	0.059 [0.039, 0.081]	−0.43 [−2.00, 1.08]
Qwen3.5-2B (thinking)	5%	128	0.06	0.17	0.114 [0.084, 0.145]	−1.37 [−3.16, 0.41]
Qwen3.5-2B (thinking)	10%	128	0.12	0.24	0.125 [0.086, 0.165]	−1.71 [−3.86, 0.45]
Qwen3.5-2B (thinking)	20%	128	0.23	0.35	0.121 [0.070, 0.175]	−1.36 [−3.85, 1.26]
Qwen3.5-2B (thinking)	40%	128	0.43	0.49	0.063 [0.002, 0.131]	−1.33 [−4.35, 1.66]

Table 7: Two candidates on one sampled prefix against independent sampling at the same expected generated tokens. Predicted: $(\rho(m) - s(m))$ times the mean of $\mu_q(1-\mu_q)$, the pooled two-candidate identity of Section 3; measured: independent minus shared coverage with the estimators of the pre-registered comparisons. Percentage points, with pointwise 95% paired bootstrap intervals; this check is not part of the analysis plan.

Model	m	$\rho - s$	Predicted	Measured
Qwen3.5-2B	8	0.13	2.6 [1.7, 3.6]	3.0 [0.3, 5.7]
Qwen3.5-2B	16	0.15	3.0 [1.9, 4.3]	3.3 [0.1, 6.7]
Qwen3.5-2B	48	0.19	3.8 [2.4, 5.2]	6.7 [2.8, 10.7]
Qwen3.5-4B	8	0.16	3.2 [2.3, 4.0]	4.2 [1.9, 6.5]
Qwen3.5-4B	16	0.20	3.8 [2.8, 4.9]	4.1 [1.4, 6.6]
Qwen3.5-4B	48	0.25	4.8 [3.4, 6.2]	5.3 [2.1, 8.7]
Qwen3.5-9B	8	0.12	2.3 [1.6, 2.9]	2.6 [1.0, 4.1]
Qwen3.5-9B	16	0.18	3.3 [2.5, 4.2]	3.1 [1.3, 5.0]
Qwen3.5-9B	48	0.20	3.7 [2.5, 5.0]	4.5 [1.9, 7.1]
OLMo-3-7B	8	0.07	1.5 [0.8, 2.3]	1.8 [−0.7, 4.2]
OLMo-3-7B	16	0.08	1.7 [0.7, 2.7]	2.4 [−0.5, 5.3]
OLMo-3-7B	48	0.06	1.3 [−0.1, 2.8]	2.2 [−1.7, 6.0]
Qwen3.5-2B (thinking)	2%	0.06	1.2 [0.8, 1.7]	1.2 [−1.1, 3.5]
Qwen3.5-2B (thinking)	5%	0.11	2.4 [1.7, 3.0]	3.8 [1.1, 6.5]
Qwen3.5-2B (thinking)	10%	0.13	2.6 [1.7, 3.4]	4.5 [1.4, 7.7]
Qwen3.5-2B (thinking)	20%	0.12	2.5 [1.4, 3.6]	3.9 [0.1, 7.5]
Qwen3.5-2B (thinking)	40%	0.06	1.3 [0.0, 2.7]	1.8 [−2.6, 6.1]

of the prefixes at any position, and among prefixes with no successful continuation, 0.2% to 9.2% of the continuations are unparsed. The allowance of 320 tokens cuts the answer of 0.8% to 1.4% of the independent samples.

Earlier records. The analysis plan compares the four-prefix records with the earlier single-prefix records on the questions they share (Table 9). The pass@1 of the new independent banks differs from the earlier banks by 0.4 to 1.3 pp. The earlier records carry one sampled prefix per question, so their estimates of $\rho(m)$ and of the 1×8 difference are less precise; the largest difference, $\rho(48)$ of Qwen3.5-9B, is 0.38 from the earlier prefixes and 0.60 from the new ones on the same 64 questions, with overlapping intervals.

Table 8: Unparsed answers in the four-prefix measurements (%): independent samples, continuations at each position, and at 48 tokens the prefixes whose continuations are mostly unparsed, the intraclass correlation of the unparsed indicator across prefixes, and the unparsed share among continuations of prefixes with no successful continuation. An unparsed answer counts as a failure in both procedures.

		Continuations			At 48 tokens		
Model	Independent	$m = 8$	16	48	Mostly unparsed	ICC	Dead prefixes
Qwen3.5-2B	2.4	2.3	2.2	2.3	0.2	0.16	3.3
Qwen3.5-4B	2.2	1.9	1.8	1.9	0.8	0.31	4.1
Qwen3.5-9B	1.3	1.7	1.5	1.5	0.8	0.35	1.0
OLMo-3-7B	11.1	10.8	11.8	11.1	3.5	0.25	9.0

Table 9: Earlier single-prefix records against the four-prefix records on the same questions (number in parentheses). Accuracy: new minus earlier pass@1 of the independent banks, percentage points with a 90% interval. $\rho(m)$ and the 1×8 difference (independent minus shared, percentage points) are computed with the same estimators from the earlier sampled prefix of each question and from the four new ones, with 95% intervals.

			$\rho(m)$		1×8 difference	
Model	Accuracy	m	Earlier	New	Earlier	New
Qwen3.5-2B (128)	1.3 [0.0, 2.6]	8	0.26 [0.18, 0.35]	0.20 [0.16, 0.25]	9.1 [5.0, 13.6]	7.0 [4.6, 9.4]
		16	0.36 [0.27, 0.46]	0.30 [0.24, 0.36]	12.7 [7.7, 18.2]	11.1 [8.1, 14.4]
		48	0.71 [0.57, 0.87]	0.64 [0.57, 0.72]	20.0 [12.5, 27.4]	25.0 [20.5, 29.4]
Qwen3.5-9B (64)	0.6 [−1.3, 2.4]	8	0.13 [0.06, 0.21]	0.21 [0.15, 0.29]	2.2 [0.0, 5.2]	3.6 [1.4, 5.9]
		16	0.25 [0.13, 0.38]	0.32 [0.23, 0.42]	6.5 [1.8, 12.5]	6.8 [3.6, 10.5]
		48	0.38 [0.26, 0.51]	0.60 [0.47, 0.75]	8.2 [2.3, 15.3]	14.7 [10.5, 19.4]
OLMo-3-7B (64)	0.4 [−1.4, 2.2]	8	0.13 [0.07, 0.21]	0.16 [0.11, 0.21]	2.3 [−1.0, 6.6]	6.7 [3.3, 10.2]
		16	0.21 [0.13, 0.32]	0.25 [0.19, 0.32]	5.6 [0.3, 11.7]	9.5 [5.8, 13.5]
		48	0.52 [0.41, 0.66]	0.56 [0.47, 0.66]	4.8 [−6.4, 16.2]	16.8 [10.9, 23.1]

Table 10: Earlier single-prefix measurements: independent minus shared pass@8 (percentage points) at the same accounted budget, $k_{\text{eq}} = 8 - 7\min(m, L_q)/L_q$, by shared tokens m. Shared coverage uses the 16-continuation bank. Bold: 95% paired bootstrap interval excludes zero.

Model	2	4	8	12	16	24	32	48	64	96
Qwen3.5-2B	**4.0**	**6.5**	**9.4**	**11.7**	**14.9**	**14.7**	**17.8**	**25.7**	**23.4**	**11.1**
Gemma-4-E4B	**5.4**	**5.7**	**14.2**	**16.1**	**19.3**	**22.5**	**28.1**	**28.4**	**25.8**	**14.4**
Qwen3.5-9B	−1.4	−0.7	−0.2	3.2	4.4	6.6	5.3	5.5	8.2	7.1
OLMo-3-7B	0.4	0.3	2.0	**6.7**	6.5	**11.1**	**12.8**	**13.1**	**13.3**	−1.3

Earlier single-prefix measurements. Table 10 reports the earlier measurements with one sampled prefix per question at ten positions. Their accounting uses $\min(m, L_q)$ in place of $\mathbb{E}[\min(T, m) \mid q]$, which is never smaller because $\min(T, m)$ is concave in T; it overstates the saving of the shared prefix and therefore favors sharing. Independent sampling covers significantly more problems in 25 of the 40 comparisons and sharing in none; four point estimates favor sharing, three of them for Qwen3.5-9B at 2 to 8 tokens.

C.1 Retries

For a prefix whose $K = 32$ continuations contain F failures, $(F)_k/(K)_k$ estimates the probability of k failures without bias and weights the prefix by its failure history; $(F)_k(K-F)/(K)_{k+1}$ estimates without bias the joint probability of k failures followed by a success, and the reported next-attempt success is the ratio of the pooled joint and history estimates. The restart arm evaluates independent samples of the same questions under identical weights. Pooled over questions, the next continuation after one failure succeeds with probability $(1-\rho)\,\mathbb{E}_q[\mu_q(1-\mu_q)]/(1-\mathbb{E}_q[\mu_q])$ by Proposition 3. The replay removes k failed continuations, chosen uniformly among the prefix's failures, as the history, orders the remaining continuations by an independent random permutation, and then continues the

Table 11: Next-attempt success (%) after k failures that come from continuations of the same prefix or from independent candidates of the same question. The last column is the restart advantage after same-prefix failures minus that after independent failures, in percentage points, with 95% paired bootstrap intervals.

			Same-prefix failures		Independent failures		
Model	m	k	Continue	Restart	Continue	Restart	Difference
Qwen3.5-2B	48	1	13.5 [11.5, 15.7]	39.6 [35.8, 43.5]	37.2 [33.0, 41.5]	39.1 [35.4, 42.9]	24.1 [20.8, 27.9]
Qwen3.5-2B	48	4	4.2 [3.4, 5.1]	39.6 [35.5, 43.9]	24.5 [20.3, 29.5]	24.9 [21.0, 29.7]	35.0 [29.6, 40.7]
Qwen3.5-4B	48	1	21.4 [18.4, 24.8]	52.5 [48.2, 56.7]	51.0 [46.3, 56.2]	51.1 [47.2, 55.3]	31.0 [26.6, 35.4]
Qwen3.5-4B	48	4	4.5 [3.4, 5.7]	51.8 [46.8, 56.8]	34.3 [28.2, 41.2]	32.5 [28.3, 38.4]	49.1 [40.7, 57.3]
Qwen3.5-9B	48	1	24.0 [21.1, 27.4]	51.9 [48.0, 55.9]	49.9 [45.3, 54.8]	51.4 [47.7, 55.2]	26.5 [22.9, 30.6]
Qwen3.5-9B	48	4	6.1 [4.9, 7.4]	49.3 [44.4, 54.4]	28.4 [23.2, 35.4]	30.9 [26.7, 36.5]	40.8 [34.5, 47.9]
OLMo-3-7B	48	1	17.6 [15.0, 20.3]	43.3 [39.8, 47.2]	40.9 [36.6, 45.6]	42.2 [39.0, 45.7]	24.4 [20.6, 28.4]
OLMo-3-7B	48	4	4.8 [3.9, 5.9]	42.9 [38.9, 47.1]	29.0 [24.2, 34.6]	29.0 [25.2, 33.6]	38.1 [31.8, 44.7]
Qwen3.5-2B (thinking)	10%	1	29.2 [25.9, 32.6]	40.3 [37.1, 43.7]	38.3 [34.7, 42.1]	40.1 [36.8, 43.6]	9.3 [7.7, 11.0]
Qwen3.5-2B (thinking)	10%	4	11.2 [9.3, 13.4]	33.8 [30.4, 37.6]	26.8 [23.4, 30.7]	27.5 [23.8, 31.9]	21.9 [18.0, 26.2]
Qwen3.5-2B (thinking)	20%	1	25.1 [22.1, 28.4]	40.2 [36.9, 43.6]	38.6 [34.7, 42.7]	40.1 [36.8, 43.6]	13.6 [11.3, 15.9]
Qwen3.5-2B (thinking)	20%	4	9.1 [7.5, 10.9]	34.4 [30.9, 38.1]	27.4 [23.2, 32.2]	27.5 [23.8, 31.9]	25.1 [20.1, 31.2]

Table 12: Replay after k failed continuations of one prefix: fraction of questions (%) solved within the additional budget of generated tokens when every further attempt continues the prefix or restarts from the question, weighted by the probability of the failure history, and their difference in percentage points, with pointwise 95% paired bootstrap intervals.

Model	m	k	Budget	Continue	Restart	Restart − continue
Qwen3.5-2B	48	1	640	33.5 [29.1, 38.3]	79.4 [73.9, 84.6]	46.0 [39.1, 52.8]
Qwen3.5-2B	48	4	640	17.9 [14.4, 21.7]	79.3 [73.5, 84.9]	61.4 [54.6, 68.2]
Qwen3.5-2B	16	1	640	60.2 [54.8, 65.9]	78.9 [73.4, 84.2]	18.7 [13.5, 24.1]
Qwen3.5-2B	16	4	640	38.2 [32.8, 44.1]	73.8 [66.4, 81.1]	35.7 [28.0, 43.1]
Qwen3.5-4B	48	1	640	40.3 [34.6, 46.5]	87.9 [84.2, 91.3]	47.6 [41.2, 54.0]
Qwen3.5-4B	48	4	640	15.6 [11.7, 19.9]	87.3 [82.9, 91.2]	71.7 [65.6, 77.3]
Qwen3.5-4B	16	1	640	65.8 [61.0, 71.1]	87.6 [84.0, 91.1]	21.8 [17.5, 25.9]
Qwen3.5-4B	16	4	640	37.2 [31.9, 43.2]	83.7 [78.5, 88.9]	46.5 [39.9, 53.0]
Qwen3.5-9B	48	1	640	46.0 [41.0, 51.5]	86.5 [83.0, 89.9]	40.5 [34.5, 46.3]
Qwen3.5-9B	48	4	640	19.8 [15.9, 24.1]	85.3 [80.8, 89.6]	65.5 [59.4, 71.5]
Qwen3.5-9B	16	1	640	67.4 [63.1, 72.0]	86.7 [83.3, 90.2]	19.3 [14.6, 24.0]
Qwen3.5-9B	16	4	640	36.8 [31.4, 42.7]	83.2 [78.3, 88.2]	46.4 [38.6, 54.0]
OLMo-3-7B	48	1	640	41.2 [36.0, 46.5]	87.1 [83.1, 90.9]	45.9 [39.1, 52.5]
OLMo-3-7B	48	4	640	19.9 [16.2, 24.0]	87.1 [82.6, 91.1]	67.2 [60.5, 73.4]
OLMo-3-7B	16	1	640	67.7 [62.5, 72.8]	86.9 [82.9, 90.7]	19.2 [14.0, 24.8]
OLMo-3-7B	16	4	640	41.9 [35.6, 48.8]	83.8 [77.8, 89.2]	41.9 [33.4, 50.4]
Qwen3.5-2B (thinking)	10%	1	24,488	52.4 [47.2, 57.9]	72.6 [68.6, 76.6]	20.2 [15.4, 25.0]
Qwen3.5-2B (thinking)	10%	4	24,488	25.4 [21.3, 30.1]	66.3 [61.2, 71.4]	40.9 [34.7, 46.4]
Qwen3.5-2B (thinking)	20%	1	24,488	46.8 [41.7, 52.2]	72.6 [68.5, 76.6]	25.8 [21.1, 30.4]
Qwen3.5-2B (thinking)	20%	4	24,488	22.4 [18.6, 26.9]	67.0 [62.0, 72.0]	44.6 [39.2, 49.6]

same prefix or restarts until 640 additional tokens are spent (for the thinking model, the tokens of four independent candidates).

Table 12 reports the replay after one and four failures at both positions. Table 11 contrasts failures of continuations of one prefix with failures of independent candidates. After independent failures, continuing a sampled prefix and restarting succeed at similar rates, so the restart advantage after shared failures comes from the prefix that the failures share.

C.2 Prediction of shared coverage

Models. Each prediction uses a held-out question's own μ_q, estimated from its independent bank, and the pooled $\rho(m)$ of Equation (8), estimated on the other half of the questions. The Beta model takes $V_m \sim \mathrm{Beta}\big(\mu_q\kappa, (1-\mu_q)\kappa\big)$ with $\kappa = 1/\rho - 1$, which has mean μ_q and variance $\rho\,\mu_q(1-\mu_q)$, so that $F_b = \prod_{j=0}^{b-1}\big((1-\mu_q)\kappa + j\big)/(\kappa + j)$. The decided mixture takes a prefix to be decided

Table 13: Mean absolute error (percentage points) of held-out predictions of shared coverage, by the number b of continuations per prefix, pooled over models and positions: $b = 2$ for 2×2 and 4×2, $b = 4$ for 1×4 and 2×4, $b = 8$ for 1×8. At $b = 2$ the three two-moment predictions coincide. The last row uses the coverage measured on the other half of the questions as the prediction.

Prediction	$b = 2$	$b = 4$	$b = 8$	All
Beta	1.3	1.7	2.4	1.7
Decided mixture	1.3	1.7	2.5	1.7
Midpoint of the bounds	1.3	1.7	2.4	1.7
Independent coverage ($\rho = 0$)	2.8	8.4	13.0	7.1
Other half, measured	2.4	2.7	2.7	2.6
Cells	34	34	17	85

Table 14: Qwen3.5-35B-A3B in the second pre-registered run, with the estimators of Tables 4 and 6: pooled profile, independent minus shared coverage at the same expected tokens and under the hard-budget replay (percentage points), and shared minus independent pass@1, with pointwise 95% paired bootstrap intervals.

				Expected budget			Hard budget	
m	$s(m)$	$\rho(m)$	$\rho - s$	1×8	2×4	4×2	1×8	pass@1
8	0.06	0.14	0.08 [0.05, 0.12]	3.5 [1.9, 5.2]	1.3 [0.1, 2.8]	0.7 [−0.5, 2.2]	3.7 [2.0, 5.6]	−0.79 [−2.37, 0.77]
16	0.11	0.24	0.13 [0.08, 0.18]	6.5 [4.5, 8.7]	2.3 [1.0, 3.9]	1.0 [−0.2, 2.6]	7.0 [4.8, 9.4]	−1.14 [−3.14, 0.80]
48	0.33	0.48	0.15 [0.07, 0.23]	14.4 [11.5, 17.5]	4.7 [2.7, 7.1]	1.5 [−0.2, 3.6]	15.0 [11.8, 18.2]	−1.90 [−4.43, 0.56]

with probability ρ, in which case V_m is one with probability μ_q and zero otherwise, and undecided otherwise, with $V_m = \mu_q$; it has the same two moments and $F_b = \rho(1 - \mu_q) + (1 - \rho)(1 - \mu_q)^b$. The midpoint prediction is the center of the coverage interval of Proposition 2. An allocation $G \times b$ is predicted to cover the question with probability $1 - F_b^G$.

Evaluation. Each of 200 random halvings of every cohort fits $\rho(m)$ on one half and predicts the mean shared coverage of the other half, which is measured with the estimator of Equation (12). Allocations with one candidate per prefix are independent sampling and are excluded, leaving 1×4, 2×2, 1×8, 2×4 and 4×2 at every position, 85 cells over the five models. At $b = 2$ the three two-moment models coincide, since F_2 depends only on the first two moments, so Section 5 reports the 51 cells with $b \in \{4, 8\}$. Table 13 reports the errors by b; the prediction with $\rho = 0$ is the coverage of independent sampling with the same number of candidates. The Beta model, the decided mixture and the midpoint of the bounds have nearly the same errors at $b \geq 4$, so the data do not favor one two-moment model over another; the Beta model overestimates shared coverage by 0.7 pp on average there, most at $b = 8$. The coverage measured on the other half predicts each cell from different questions, without the held-out questions' own μ_q.

D Second Pre-Registered Run

A second analysis plan, fixed before its sampling, added Qwen3.5-35B-A3B to the four-prefix measurements, measured prefix selection with every sampled prefix charged, and planned a thinking model on competition problems (Appendix A.1). **A larger base model.** Qwen3.5-35B-A3B repeats the four-prefix measurements of Section 2 with the same estimators (Table 14). Its profile lies significantly above the line at all three positions, with ratios of 1.5 to 2.5, and eight candidates sharing one prefix cover significantly fewer problems than independent sampling at every position, under the expected and the hard budget; no allocation covers significantly more. After four failed continuations of a 48-token prefix, a further continuation succeeds in 6.3% of attempts and a restart in 49.0%, and the Beta prediction of Section 5 has a mean absolute error of 1.5 pp, against 2.2 pp for the coverage measured on the other half.

Prefix selection. From $G \in \{2, 4\}$ of the four sampled prefixes of a question, a score fixed before sampling keeps one prefix, which is continued eight times; the cost counts all G prefixes, $G\,\mathbb{E}[\min(T, m)] + 8\ell_m$ generated tokens, where ℓ_m is the mean useful length of the kept prefix's

Table 15: Prefix selection with Qwen3.5-35B-A3B on GSM8K. From G of the four sampled prefixes, the one with the highest score is kept and continued eight times; every sampled prefix is charged. Columns: independent minus selected coverage at the same expected generated tokens, selected minus randomly kept coverage within the same subsets, and the change in pass@1 of the kept prefix's continuations, in percentage points with pointwise 95% paired bootstrap intervals. At $m \leq 32$ the last-32 confidence equals the prefix confidence and is omitted.

m	G	Score	Independent − selected	Selected − random	pass@1 change
8	2	Log-probability	2.9 [1.2, 5.0]	0.6 [−0.5, 1.7]	0.8 [−1.0, 2.6]
8	2	Confidence	3.1 [1.4, 5.1]	0.4 [−0.4, 1.2]	0.0 [−1.9, 1.9]
8	4	Log-probability	3.5 [1.1, 6.5]	0.0 [−2.1, 1.9]	1.5 [−1.2, 4.2]
8	4	Confidence	3.0 [0.8, 5.7]	0.3 [−1.4, 1.8]	−0.5 [−3.5, 2.7]
16	2	Log-probability	5.3 [3.2, 7.7]	1.2 [0.0, 2.6]	0.1 [−2.3, 2.2]
16	2	Confidence	6.3 [3.8, 9.0]	0.3 [−1.1, 1.8]	−1.0 [−3.6, 1.6]
16	4	Log-probability	5.4 [2.5, 8.7]	1.2 [−1.5, 3.7]	0.6 [−2.9, 4.0]
16	4	Confidence	8.0 [4.1, 12.2]	−1.4 [−4.7, 1.8]	−2.8 [−7.1, 1.4]
48	2	Log-probability	12.9 [9.4, 16.5]	1.9 [−0.2, 4.0]	−0.3 [−3.7, 2.7]
48	2	Confidence	13.4 [9.9, 17.2]	1.2 [−1.0, 3.5]	−1.0 [−4.2, 2.2]
48	2	Last-32 confidence	12.1 [8.7, 15.8]	2.5 [0.3, 4.8]	0.4 [−2.8, 3.5]
48	4	Log-probability	12.3 [7.3, 17.7]	2.6 [−1.9, 6.8]	2.2 [−2.9, 7.0]
48	4	Confidence	12.0 [7.1, 17.5]	2.7 [−1.7, 7.1]	1.9 [−3.7, 7.1]
48	4	Last-32 confidence	12.6 [7.8, 18.1]	2.2 [−2.3, 6.1]	1.5 [−4.0, 6.6]

Table 16: Majority-vote accuracy (%) of Qwen3.5-35B-A3B at the same expected generated tokens: independent sampling at k_{eq} candidates against shared allocations, with the difference in percentage points and its pointwise 95% paired bootstrap interval. Unparsed answers abstain and ties are split uniformly.

m	Allocation	Independent	Shared	Difference
8	1×8	85.7	80.9	4.8 [2.8, 6.9]
8	2×4	85.7	84.2	1.5 [−0.3, 3.3]
8	4×2	85.9	85.9	0.0 [−1.7, 1.8]
16	1×8	85.3	77.6	7.7 [5.2, 10.3]
16	2×4	85.4	81.8	3.7 [1.4, 5.9]
16	4×2	85.6	84.6	1.1 [−1.0, 3.1]
48	1×8	83.4	72.6	10.9 [8.0, 13.8]
48	2×4	84.0	77.4	6.5 [3.6, 9.6]
48	4×2	84.8	83.2	1.7 [−1.3, 4.7]

continuations. The scores are the mean log-probability of the sampled tokens and the mean top-20 confidence, the negative mean log-probability of the 20 most likely tokens at each position (Fu et al., 2026), over the whole prefix or, for the confidence, over its last 32 tokens; ties keep the prefix sampled first. The coverage of eight continuations of the kept prefix is $1 - (F)_8/(32)_8$ from its 32 continuations, averaged over all G-subsets; keeping a random prefix of the subset gives the coverage of 1×8 at the higher cost. In all 14 settings of Table 15, independent sampling at the same expected tokens covers significantly more problems; the score-selected prefix beats random choice significantly in one, and the pass@1 of its continuations changes by -2.8 to $2.2\,\text{pp}$, with every interval containing zero.

Majority vote, execution order and retries. With the same candidates, majority voting, in which unparsed answers abstain and ties are split at random, also favors independent sampling at the same expected tokens, by 4.8 to 10.9 pp for 1×8 at the three positions (Table 16). Under the hard-budget replay, completing each prefix's continuations before starting the next prefix changes coverage by at most 0.4 pp from drawing continuations in turn across the prefixes. In the retry replay with 640 additional tokens, continuing a failed prefix once before restarting solves 1.4 to 3.6 pp fewer questions than restarting after every failure.

Competition problems. The plan called for measuring Qwen3.5-9B in its reasoning mode (revision `c202236`) on 265 integer-answer problems from the competitions collected by MathArena

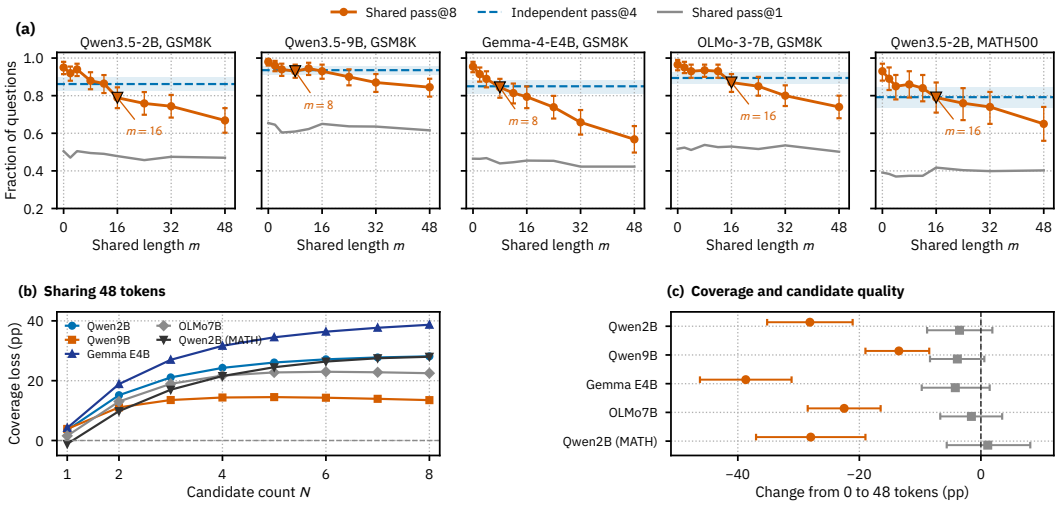

Figure 7: Five-setting scan with fresh allowances. (a) Eight-candidate coverage by sharing length; triangles mark the crossing below independent pass@4. (b) Coverage loss at 48 shared tokens for $N \leq 8$. (c) Paired changes in pass@8 and pass@1. Bars and bands: 95% bootstrap intervals.

(Balunović et al., 2025) (AIME 2023 to 2026, HMMT, BRUMO, SMT and CMIMC), under the matched policy of the thinking model with an allowance of 32,768 tokens, on problems with one to seven correct of eight screening samples and at most one screening sample reaching the allowance. Only 10 problems met this rule, below the pre-registered minimum of 48, so the profile of this model, and the branching triggered by low confidence that depended on it, were not collected. In screening, 48% of the 2,120 samples reached the allowance, 964 of them without closing the reasoning block, and 1,087 of the 1,098 samples that stopped before it were correct.

E Broader Measurements: Fresh Allowances, Other Tasks, and Separate Batches

This appendix tests whether the equal-budget ordering of Section 2 and the early resolution of success variance in Section 4 persist under an earlier design, which restarts the generation allowance after the prefix or continues one sampled prefix per question, and whether they extend to competition mathematics and to Chinese grade-school problems.

E.1 Five settings with fresh allowances

Setup. The scan covers Qwen3.5-2B, Qwen3.5-9B, Gemma-4-E4B and OLMo-3-7B on GSM8K, and Qwen3.5-2B on MATH500 (Hendrycks et al., 2021; Lightman et al., 2024). Questions pass the intermediate-difficulty screen and have eight answers at every measured length: 199 to 200 per grade-school setting and 100 on MATH500. Sharing lengths are $m \in \{0, 2, 4, 8, 12, 16, 24, 32, 48\}$. Each continuation receives a fresh allowance of 320 tokens on GSM8K and 640 on MATH500, whereas the matched policy of Section 2 continues within the allowance that remains after the prefix. OLMo-3-7B's samples in this scan continue after their answer with unrelated text that the `Question:` marker does not cut, and grading, which takes the last answer in the `####` format before that marker, reads a different answer from this text for 1.5% of its independent samples.

Results. For Qwen3.5-2B on GSM8K, eight independent candidates raise coverage from 50.5% to 95.0%, whereas eight candidates sharing their first 48 tokens reach 66.8%, below the 86.2% of four independent candidates (Figure 7). Across the five settings, 48 shared tokens lower pass@8 by 13.5 to 38.7 pp, and in every setting eight shared candidates cover fewer problems than four independent ones, with paired 95% intervals excluding zero. At 48 tokens, pass@1 changes by -4.2 to $+1.1$ pp and all five intervals contain zero, so the loss arises from the dependence among candidates. Sharing removes 30 to 70% of the coverage gain from seven additional candidates (Appendix B.1). With a

Table 17: Coverage (%) at 48 shared tokens; paired differences in percentage points with 95% CIs.

(a) Fresh continuation allowances: coverage (%) and gain removed

Model	Task	n	Ind.@4	Ind.@8	Shr.@8	Gain removed (%)	Ind.@4 − Shr.@8 [95% CI]
Qwen3.5-2B	GSM8K	199	86.18	94.97	66.83	55.4	**19.35** [12.79, 25.97]
Qwen3.5-2B	MATH500	100	79.20	93.00	65.00	54.1	**14.20** [5.50, 23.00]
Qwen3.5-9B	GSM8K	200	93.56	98.00	84.50	29.6	**9.06** [4.48, 13.87]
Gemma-4-E4B	GSM8K	199	84.99	95.48	56.78	70.4	**28.21** [21.08, 35.43]
OLMo-3-7B	GSM8K	200	89.41	96.50	74.00	46.8	**15.41** [9.48, 21.50]

(b) Single-prefix measurements under the matched policy, GSM8K

Model	n	Ind.@4	Ind.@8	Shr.@8	Ind.@4 − Shr.@8 [95% CI]
Qwen3.5-2B	128	83.54	93.64	62.10	**21.44** [13.63, 29.02]
Gemma-4-E4B	128	86.95	95.52	60.45	**26.51** [18.19, 35.04]
Qwen3.5-9B	64	91.42	95.73	87.89	**3.53** [−4.56, 11.95]
OLMo-3-7B	64	88.58	96.72	74.99	**13.59** [4.00, 23.47]

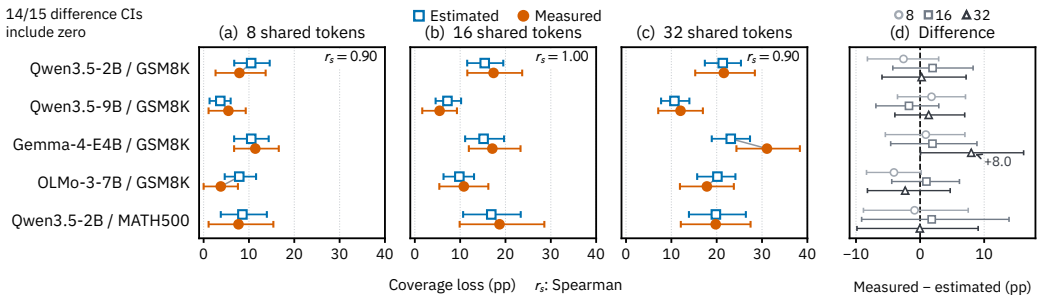

Figure 8: (a–c) Estimated and separately measured eight-candidate coverage losses; (d) measured minus estimated. Bars: 95% question-bootstrap intervals.

common estimated answer length L per setting, eight candidates sharing 48 tokens cost as much as $k_{\text{eq}} = 8 - 7 \cdot 48/L$ independent candidates, 4.3 to 5.6 for L from 91 to 139 tokens; this accounting overstates the saving, since $\mathbb{E}[\min(T, 48)] \leq \min(48, L)$, and so favors sharing. On the 91 to 195 questions per setting with length measurements, independent $\text{pass@}\lfloor k_{\text{eq}} \rfloor$ exceeds shared pass@8 by 11.3 to 28.2 pp, with all five paired intervals excluding zero (Table 19).

E.2 Coverage from prefix measurements in a separate batch

The coverage cost of Theorem 1 can be estimated from conditional-success measurements and compared with coverage measured in a separate batch of continuations. At $m = 8, 16, 32$, one randomly selected correct source and one parsed incorrect source per question are continued eight times with a fresh allowance, and their falling-factorial estimates are combined with the independent-sample success fraction. Across fifteen setting–length combinations, estimated and separately measured eight-candidate losses differ by 2.0 pp in mean absolute error, with Pearson correlation 0.94; fourteen of the fifteen measured-minus-estimated intervals contain zero (Figure 8). Holding the sharing length fixed, the estimates recover the ordering of the five settings with Spearman correlations of 0.90, 1.00 and 0.90 at 8, 16 and 32 tokens. The same shared length thus costs different amounts of coverage in different settings, and conditional-success measurements account for which settings pay more.

Interval measurements. In the single-prefix measurements, separate banks of 64 and 16 continuations from the same source paths estimate squared changes and coverage losses over ten adjacent intervals from 0 to 96 tokens. Mean absolute errors range from 1.0 to 1.7 pp. Cumulative estimates

Table 18: Halfway positions of $r(m)$ by model and task (tokens).

Model	GSM8K	MATH500	MGSM-zh
Qwen3.5-2B	34.8	48.6	32.0
Qwen3.5-9B	49.0	> 64	43.7
Gemma-4-E4B	36.9	45.0	—
OLMo-3-7B	41.9	—	50.4

are 18.5 pp against 18.8 pp measured for Qwen3.5-2B and 19.8 pp against 13.1 pp for OLMo-3-7B; Appendix F.3 reports all four models.

E.3 Profiles across models and tasks

For sixteen model–task settings, we measure continuation success from correct source paths with fresh allowances. Their normalized profile $r(m)$ measures how far correct-source continuation success rises above the success of independent samples and equals $\rho(m)$ when the continuation policy is matched (Appendix F.5). Table 18 gives the first positions where $r(m)$ reaches one half, interpolated on the logarithmic token grid. Across ten GSM8K models these positions range from 32 to 49 tokens; on MGSM-zh (Shi et al., 2023), they range from 32 to 50. MATH500 reaches one half at 45 tokens or later, with one profile still below one half at 64 tokens. For Qwen3.5-2B, the position shifts from 32 tokens on MGSM-zh to 48.6 on MATH500. Longer and harder solutions move the resolution of success variance later in tokens, while it still occurs within the first few dozen tokens.

F Detailed Coverage, Prediction, and Profile Results

This appendix gives the cohort definitions, estimators and complete tables behind the broader measurements of Appendix E, the single-prefix measurements that Sections 2 and 4 cite, and the greedy-prefix measurements of Appendix I.

F.1 Coverage and generated-token budgets

Table 17(a) in Appendix E.1 reports the sharing-length comparison of the scan. Independent and shared coverage use the same questions complete across sharing lengths; n gives that paired count. "Gain removed" is one minus the ratio of the shared gain to the independent gain, with each gain measured from that arm's own pass@1, as defined in Appendix B.1. The final column compares four independent candidates with eight candidates sharing 48 tokens.

The scan samples one source uniformly from the question's sixteen independent samples, without conditioning on its correctness. The independent bank is the first eight samples from that same pool. Each source prefix is decoded to text and re-encoded with the prompt before eight new continuations are sampled at temperature 1.0 and top-$p = 0.95$. The generation allowance is restarted at each prefix. A source shorter than a requested position supplies no continuation at that position, and the paired cohort retains questions with all eight answers at every position. Figure 7(b) computes coverage for $N = 1, \ldots, 8$ by averaging $1 - \binom{8-k}{N} / \binom{8}{N}$ over questions with k correct answers in the relevant bank. Its endpoint differences and the paired intervals in panel (c) use the same questions as panel (a).

Table 19 reports the separate cohort with length measurements for the generated-token comparison. Here L is the mean answer length, k_{eq} is given by Equation (2), and $k = \lfloor k_{\text{eq}} \rfloor$. The final column is independent pass@k minus shared pass@8; the interval is the paired 95% bootstrap interval on that difference.

The budget model uses a common mean answer length L for both procedures. For each stored independent sample, the length analysis truncates text at the end of its first answer or at the next `Question:` marker, whichever comes first, which excludes the text after OLMo-3-7B's answers (Appendix E.1), and multiplies its recorded token count by the fraction of characters retained. Averaging these estimated answer lengths gives L. Independent coverage at $k = \lfloor k_{\text{eq}} \rfloor$ is estimated from sixteen independent answers on the budget cohort.

Table 19: Generated-token accounting with a common mean answer length. Differences include paired 95% bootstrap intervals.

Model	Task	n	L (tokens)	k_{eq}	Ind.@k (%)	Shr.@8 (%)	Difference (pp) 95% CI
Qwen3.5-2B	GSM8K	192	100.9	4.67	86.9	66.7	20.2 [13.8, 26.7]
Qwen3.5-2B	MATH500	91	123.9	5.29	87.0	68.1	18.9 [9.9, 27.7]
Qwen3.5-9B	GSM8K	183	139.0	5.58	95.5	84.2	11.3 [6.5, 16.4]
Gemma-4-E4B	GSM8K	195	90.6	4.29	85.1	56.9	28.2 [21.0, 35.5]
OLMo-3-7B	GSM8K	185	90.9	4.30	88.7	73.5	15.2 [9.2, 21.4]

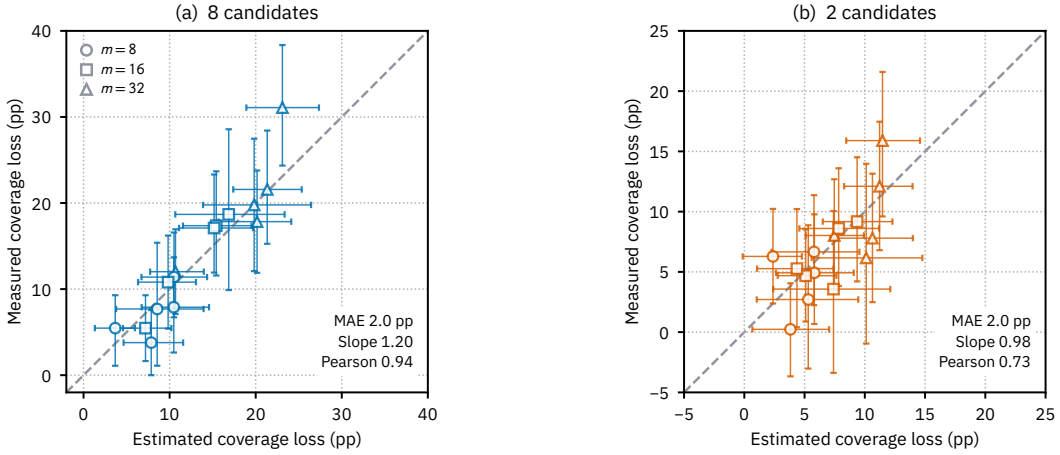

Figure 9: Estimated and separately measured coverage losses for (a) eight and (b) two candidates. Each panel contains five settings at three sharing lengths. Bars show 95% question-bootstrap intervals; the dashed line marks equality.

F.2 Eight-candidate estimates at every measured length

Table 20 gives all fifteen setting–length pairs in Figure 8, including their cohort sizes and uncertainty. The measured loss and its difference from the estimate each have a 95% question-bootstrap interval. Differences are measured minus estimated and are computed before rounding. Gemma-4-E4B at 32 tokens is the one difference interval excluding zero.

For each question, the measurement batch contains one correct and one parsed incorrect source path, each sampled uniformly from its stratum in the pool of sixteen independent samples. Their failure estimates are combined with weights $\widehat{\mu}$ and $1-\widehat{\mu}$, where $\widehat{\mu}$ is the independent-sample success fraction. The latter weight includes all independent-sample failures, while the incorrect-source stratum requires a parsed answer. Equation (12) also estimates the independent failure probability from all sixteen independent samples; the scan's observed independent coverage uses their first eight. Source prefixes are decoded and re-encoded, and each continuation starts with a fresh allowance of 320 tokens on GSM8K or 640 on MATH500. Questions are retained when both source types have continuation measurements at 8, 16 and 32 tokens and the scan has coverage measurements at the same positions. The bootstrap uses 1,000 resamples of questions within each setting.

Figure 9 compares estimates and separate measurements for eight and two candidates. For eight candidates, regressing measured loss on estimated loss with an intercept gives a slope of 1.20 and an intercept of -2.1 pp. Comparing the shared-prefix failure probabilities $\mathbb{E}[(1-v)^8]$ directly, with v defined by the continuation protocol as in Appendix E.2, gives a mean absolute error of 2.2 pp, a slope of 1.10 and a Pearson correlation of 0.93. For two candidates, the fifteen loss estimates have a mean absolute error of 2.0 pp, a slope of 0.98 and a Pearson correlation of 0.73. Their fixed-length Spearman correlations are 0.20, 0.60 and 0.70 at 8, 16 and 32 tokens, respectively.

Table 20: Per-setting pass@8 estimates and measurements. Losses and differences are in percentage points; brackets give 95% bootstrap intervals.

Model	Task	m	n	Estimated	Measured [95% CI]	Difference [95% CI]
Qwen3.5-2B	GSM8K	8	190	10.5	7.9 [2.6, 13.7]	−2.6 [−8.2, 2.9]
Qwen3.5-2B	GSM8K	16	190	15.4	17.4 [11.6, 23.7]	+1.9 [−4.3, 8.3]
Qwen3.5-2B	GSM8K	32	190	21.3	21.6 [15.3, 28.4]	+0.2 [−5.9, 7.2]
Qwen3.5-2B	MATH500	8	91	8.6	7.7 [1.1, 15.4]	−0.9 [−8.8, 7.5]
Qwen3.5-2B	MATH500	16	91	16.9	18.7 [9.9, 28.6]	+1.8 [−9.1, 13.9]
Qwen3.5-2B	MATH500	32	91	19.8	19.8 [12.1, 27.5]	−0.0 [−9.8, 9.1]
Qwen3.5-9B	GSM8K	8	183	3.7	5.5 [1.1, 9.3]	+1.8 [−3.5, 7.1]
Qwen3.5-9B	GSM8K	16	183	7.2	5.5 [1.6, 9.3]	−1.7 [−6.9, 2.9]
Qwen3.5-9B	GSM8K	32	183	10.7	12.0 [7.1, 17.0]	+1.3 [−3.9, 7.0]
Gemma-4-E4B	GSM8K	8	193	10.5	11.4 [6.7, 16.6]	+0.9 [−5.4, 7.0]
Gemma-4-E4B	GSM8K	16	193	15.2	17.1 [11.9, 23.3]	+1.9 [−4.6, 8.9]
Gemma-4-E4B	GSM8K	32	193	23.1	31.1 [24.4, 38.4]	+8.0 [+0.1, 16.2]
OLMo-3-7B	GSM8K	8	185	7.9	3.8 [0.0, 7.6]	−4.1 [−8.4, 0.2]
OLMo-3-7B	GSM8K	16	185	9.8	10.8 [5.4, 16.2]	+1.0 [−4.4, 6.1]
OLMo-3-7B	GSM8K	32	185	20.2	17.8 [11.9, 23.8]	−2.3 [−8.2, 4.7]

Table 21: Single-prefix measurements: predicted and measured two-candidate interval losses. Losses and MAE are in percentage points; brackets give measured 95% intervals.

							Cumulative loss, 0–96	
Model	n	Intervals	MAE	Slope	Spearman	Within CI	Predicted	Measured [95% CI]
Qwen3.5-2B	128	10	1.5	1.02	0.58	9 / 10	18.5	18.8 [11.2, 26.2]
Gemma-4-E4B	128	10	1.7	1.32	0.26	10 / 10	18.3	22.5 [14.4, 31.1]
Qwen3.5-9B	64	10	1.0	1.18	0.45	10 / 10	12.0	8.4 [−0.4, 18.3]
OLMo-3-7B	64	10	1.7	1.36	0.49	10 / 10	19.8	13.1 [2.6, 24.1]

F.3 Single-prefix measurements: two-candidate intervals

Table 21 reports the single-prefix measurements on GSM8K: 128 questions each for Qwen3.5-2B and Gemma-4-E4B, and 64 each for Qwen3.5-9B and OLMo-3-7B. Questions follow the intermediate-difficulty screen of Section 2. One sampled source path per question supplies the exact generated token IDs at positions $0, 2, 4, 8, 12, 16, 24, 32, 48, 64, 96$. Continuations follow the matched policy of Section 2: the decoding settings of the source and the remaining $320 - m$ tokens. Separate banks of 64 and 16 continuations estimate the squared change in conditional success probability and the two-candidate coverage loss, respectively; terminated paths keep their outcomes.

The MAE, slope and Spearman correlation use the ten adjacent intervals. The slope regresses measured loss on predicted loss with an intercept; Spearman measures agreement in the ordering of interval costs. "Within CI" counts predicted point estimates inside the measured 95% interval: nine of ten for Qwen3.5-2B, with the 8–12 interval outside, and ten of ten for each other model. Cumulative predictions and measurements cover the full 0–96 interval; their uncertainty uses 4,000 question-bootstrap resamples. The cumulative predictions are 18.5, 18.3, 12.0 and 19.8 pp for Qwen3.5-2B, Gemma-4-E4B, Qwen3.5-9B and OLMo-3-7B, against measured losses of 18.8, 22.5, 8.4 and 13.1 pp.

The same evaluation bank gives the candidate-count comparison in Table 17(b). Table 22 reports changes in pass@1 and eight-candidate coverage. At 48 tokens, OLMo-3-7B's pass@1 rises by 9.2 pp with a 95% interval of $[0.6, 17.7]$; the other three pass@1 intervals contain zero. Each question contributes one sampled source, shared by both continuation banks, so question resampling includes variation across these source draws along with continuation sampling.

Table 22: Single-prefix measurements: changes in pass@1 and pass@8 at 48 shared tokens. Entries are percentage points with paired 95% question-bootstrap intervals.

Model	n	Shr.@1 − Ind.@1	Ind.@8 − Shr.@8
Qwen3.5-2B	128	−1.56 [−8.40, 5.57]	31.55 [23.71, 39.24]
Gemma-4-E4B	128	−4.25 [−11.52, 2.64]	35.08 [26.89, 43.31]
Qwen3.5-9B	64	5.08 [−2.44, 12.21]	7.85 [−0.46, 16.78]
OLMo-3-7B	64	9.18 [0.59, 17.68]	21.72 [11.89, 31.91]

Table 23: Gemma-4-E4B cumulative two-candidate losses from token 0; $n = 128$. Values are percentage points with 95% intervals.

Endpoint t	Predicted	Measured	Measured − predicted
2	2.9 [1.9, 4.1]	6.2 [1.9, 10.5]	3.3 [−0.7, 7.4]
16	7.6 [6.0, 9.3]	12.4 [6.1, 19.2]	4.9 [−1.4, 11.1]
32	11.6 [9.3, 13.9]	18.5 [11.3, 26.0]	6.9 [−0.1, 13.9]
64	15.8 [13.3, 18.3]	21.8 [14.2, 29.9]	6.1 [−1.7, 14.0]
96	18.3 [15.7, 20.9]	22.5 [14.4, 31.1]	4.2 [−4.7, 13.3]

For Gemma-4-E4B, the interval predictions have a Spearman correlation of 0.26 with the measured costs, and Table 23 follows the cumulative comparison from token 0 to every recorded endpoint: the point predictions lie below the measurements at all five endpoints, by 3.3 to 6.9 pp, and every paired interval contains zero.

F.4 Single-prefix profiles and prefix states

Let S_{qm} be the number of successes in the 64-continuation measurement bank for question q at position m, and write $K = 64$. For $m > 0$, define

$$\widehat{d}_{qm} = \frac{(S_{qm})_2}{(K)_2} + \frac{(S_{q0})_2}{(K)_2} - 2\frac{S_{qm}S_{q0}}{K^2}, \qquad \widehat{b}_q = \frac{S_{q0}}{K} - \frac{(S_{q0})_2}{(K)_2}. \tag{13}$$

Because the bank at $m = 0$ is independent of the source prefix, $\widehat{d}_{qm}$ is unbiased for $(V_m - \mu_q)^2$ given the prefix, $\mathbb{E}[\widehat{d}_{qm} \mid \mathcal{F}_m] = (V_m - \mu_q)^2$, and $\widehat{b}_q$ is unbiased for $\mu_q(1 - \mu_q)$; averaging the first over sources drawn from π gives $\mathrm{Var}(V_m \mid q)$. The profile estimate $\widehat{\rho}(m) = \sum_q \widehat{d}_{qm} / \sum_q \widehat{b}_q$ is a ratio estimator: it is defined when the denominator is positive, it is not unbiased in general, and it is set to zero at $m = 0$.

The prefix state is determined from the source's exact tokens at each measured position. A terminated path has reached EOS or exhausted its generation allowance. For an unfinished path, we decode its first m tokens and check for the literal `####` before any subsequent `Question:` prompt. For state g, its normalized contribution is

$$\widehat{\rho}_g(m) = \frac{\sum_q \mathbf{1}\{G_{qm} = g\}\widehat{d}_{qm}}{\sum_q \widehat{b}_q}, \qquad \sum_g \widehat{\rho}_g(m) = \widehat{\rho}(m). \tag{14}$$

The state can change along a source path. The colored areas in Figure 10 partition each position's total by the states attained there. All 384 model–question observations and all eleven positions are retained. Intervals use 4,000 resamples of questions, including their sources and both continuation banks, and are pointwise. Estimates and intervals are reported without clipping to $[0, 1]$. Unparsed answers count as failures. Among the continuations generated from unfinished prefixes, they make up 0.8 to 9.1% at every position from 2 to 64 tokens, within 1.3 points of each model's share among its independent samples, and at 48 tokens none of the 83 unfinished prefixes without a successful continuation has a majority of unparsed continuations.

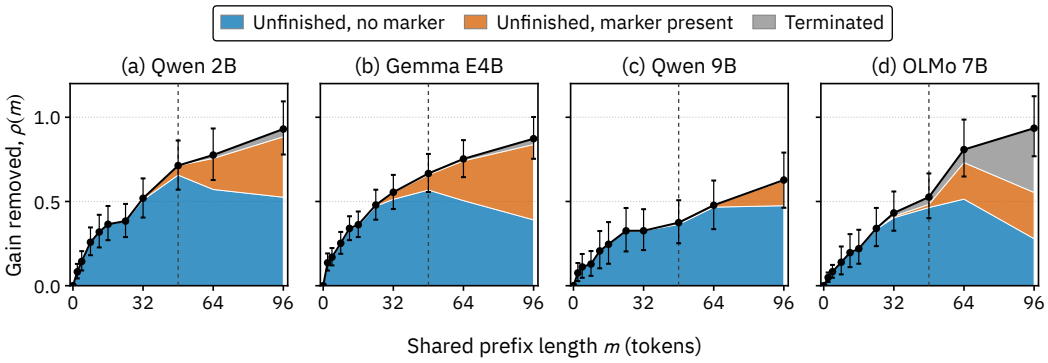

Figure 10: Single-prefix profiles $\rho(m)$ by model. Points and bars: estimates and pointwise 95% intervals. Stacks partition each estimate by the state of the source prefix at that position; the answer marker is ####. Dashed lines mark 48 tokens.

Table 24: Single-prefix profiles and prefix states at 32, 48, and 96 tokens. U: unfinished without the answer marker; A: unfinished with it; T: terminated. Counts partition each model's sources. The last column is U's share of the total estimated squared change, in percent. Brackets give pointwise 95% intervals.

m	Model	$\rho(m)$	n_U	n_A	n_T	U share (%)
32	Qwen3.5-2B	0.519 [0.405, 0.637]	126	2	0	98.2 [94.6, 100.0]
32	Gemma-4-E4B	0.555 [0.455, 0.658]	123	5	0	92.4 [83.8, 99.0]
32	Qwen3.5-9B	0.326 [0.212, 0.454]	64	0	0	100.0 [100.0, 100.0]
32	OLMo-3-7B	0.432 [0.326, 0.559]	60	2	2	93.4 [84.0, 99.5]
48	Qwen3.5-2B	0.713 [0.570, 0.862]	122	6	0	92.0 [83.4, 99.2]
48	Gemma-4-E4B	0.666 [0.556, 0.782]	117	10	1	85.2 [75.2, 94.1]
48	Qwen3.5-9B	0.374 [0.252, 0.507]	63	1	0	97.3 [90.5, 100.0]
48	OLMo-3-7B	0.526 [0.401, 0.666]	52	8	4	88.2 [77.4, 96.2]
96	Qwen3.5-2B	0.930 [0.779, 1.094]	81	36	11	56.5 [44.6, 68.8]
96	Gemma-4-E4B	0.873 [0.753, 1.002]	61	62	5	44.8 [34.1, 56.1]
96	Qwen3.5-9B	0.627 [0.462, 0.790]	50	12	2	75.7 [58.9, 89.3]
96	OLMo-3-7B	0.935 [0.768, 1.125]	21	20	23	29.9 [16.5, 44.9]

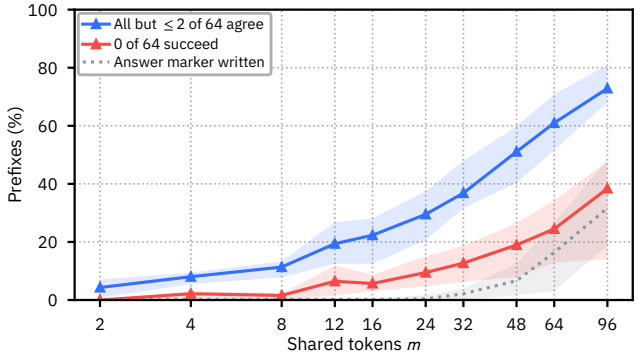

Figure 11: States of single prefixes that have neither written the answer marker nor stopped: fraction whose 64 continuations agree in all but at most two outcomes, and fraction with no successful continuation; dotted: fraction of all prefixes that contain the marker. Lines: mean over the four models; bands: range across them.

Table 25: Matched profile $\rho(m)$ and correct-source profile $r(m)$ on GSM8K, shown as $\rho(m)$ / $r(m)$: upper block, the single-prefix measurements of four base models; lower block, the four-prefix measurements of five base models at their shared lengths, 8 and 16 tokens. n gives the questions of the two cohorts, which are separate.

Model	n	$m = 2$	4	8	16	32	64
Qwen3.5-2B	128/243	0.08 / 0.02	0.14 / 0.09	0.26 / 0.19	0.36 / 0.30	0.52 / 0.48	0.78 / 0.68
Qwen3.5-9B	64/278	0.08 / 0.08	0.11 / 0.10	0.13 / 0.24	0.25 / 0.28	0.33 / 0.37	0.48 / 0.58
Gemma-4-E4B	128/252	0.14 / 0.09	0.17 / 0.12	0.25 / 0.15	0.36 / 0.32	0.55 / 0.42	0.75 / 0.80
OLMo-3-7B	64/282	0.05 / 0.10	0.08 / 0.13	0.14 / 0.13	0.22 / 0.25	0.43 / 0.43	0.81 / 0.61
Four-prefix measurements, $m = 8$ and 16							
Qwen3.5-2B	128/243			0.20 / 0.19	0.30 / 0.30		
Qwen3.5-4B	128/283			0.23 / 0.21	0.32 / 0.30		
Qwen3.5-9B	192/278			0.18 / 0.24	0.29 / 0.28		
Qwen3.5-35B-A3B	128/137			0.14 / 0.28	0.24 / 0.37		
OLMo-3-7B	128/282			0.16 / 0.13	0.27 / 0.25		

F.5 Correct-source profiles for sixteen settings

Figure 4(b) and Tables 26 and 27 report the sixteen correct-source profiles. Let W_m be the success probability under the fresh-allowance continuation protocol, and write $a_{C,q}(m) = \mathbb{E}[W_m \mid Z = 1, q]$. The normalized profile is

$$r(m) = \frac{\sum_q \{\mu_q a_{C,q}(m) - \mu_q^2\}}{\sum_q \mu_q(1 - \mu_q)}. \tag{15}$$

Because W_m is a function of the source prefix, $\mu_q a_{C,q}(m) = \mathbb{E}[V_m W_m \mid q]$. Consequently,

$$\mu_q a_{C,q}(m) - \mu_q^2 = \mathrm{Var}(V_m \mid q) + \mathbb{E}[V_m(W_m - V_m) \mid q]. \tag{16}$$

Matching the continuation policy sets $W_m = V_m$ and gives $r(m) = \rho(m)$. With fresh allowances, $r(m)$ also includes the effect of that continuation protocol. Table 25 compares the two profiles wherever a model has both. Their cohorts are separate, so each difference combines the protocol term above with differences between cohorts. Against the single-prefix measurements of four base models, the profiles differ through 32 tokens by 0.05 on average and by at most 0.13, and at 64 tokens by at most 0.20; against the four-prefix measurements of five base models, at 8 and 16 tokens, by 0.04 on average and by at most 0.15, for Qwen3.5-35B-A3B, whose correct-source profile lies above its matched one at both lengths. At every length the 95% intervals overlap.

Cohorts contain questions with a valid correct source path at every measured position: six positions on GSM8K and four on MATH500 and MGSM-zh. Write k_0 for the number of successes among sixteen independent samples, $\widehat{\mu} = k_0/16$, and $\widehat{H}_0 = k_0(k_0 - 1)/(16 \cdot 15)$. If $\widehat{a}_C(m)$ is the success fraction from the correct source prefix, the recorded estimator sums $\widehat{\mu}\widehat{a}_C(m) - \widehat{H}_0$ over questions and divides by the sum of $\widehat{\mu} - \widehat{H}_0$. All values use the same complete cohort within each setting.

Profile intervals are 95% question-bootstrap percentiles from 1,000 replicates. The tables retain negative and nonmonotone estimates. Halfway positions use interpolation on the logarithmic token grid. Their displayed intervals are conditional on a bootstrap replicate reaching one half by 64 tokens; u reports the percentage of replicates without such a crossing. The symbol > 64 marks a point estimate that does not reach one half in the measured range. At 32 tokens, the MATH500 profiles lie between 0.295 and 0.349, compared with 0.368 to 0.498 across the ten GSM8K models. The ten GSM8K halfway positions range from 32.5 (Qwen3.5-35B-A3B) to 49.0 (Qwen3.5-9B); Gemma-4-31B reaches halfway at 42.4 tokens.

Table 26: GSM8K correct-source profiles, halfway positions (tokens), and recorded 95% intervals.

Model	n	$r(2)$	$r(4)$	$r(8)$	$r(16)$	$r(32)$	$r(64)$	$m_{1/2}$	u
Qwen3.5-2B	243	0.021	0.085	0.186	0.304	0.475	0.679	34.8	0.0
		[−0.043, 0.089]	[0.015, 0.157]	[0.119, 0.253]	[0.227, 0.376]	[0.401, 0.543]	[0.617, 0.740]	[27.3, 42.1]	
Qwen3.5-4B	283	0.104	0.112	0.215	0.299	0.430	0.692	38.5	0.0
		[0.034, 0.171]	[0.029, 0.187]	[0.141, 0.280]	[0.225, 0.367]	[0.363, 0.494]	[0.633, 0.745]	[32.6, 44.4]	
Qwen3.5-9B	278	0.080	0.103	0.244	0.285	0.368	0.583	49.0	2.1
		[−0.002, 0.156]	[0.034, 0.177]	[0.178, 0.308]	[0.204, 0.356]	[0.285, 0.450]	[0.502, 0.648]	[38.8, 60.7]	
Qwen3.5-35B-A3B	137	0.089	0.164	0.284	0.366	0.498	0.592	32.5	4.4
		[−0.035, 0.192]	[0.052, 0.277]	[0.175, 0.384]	[0.251, 0.472]	[0.398, 0.592]	[0.463, 0.711]	[21.4, 54.0]	
Gemma-4-E2B	268	0.073	0.139	0.174	0.297	0.416	0.721	38.7	0.0
		[0.016, 0.125]	[0.086, 0.194]	[0.119, 0.229]	[0.237, 0.354]	[0.347, 0.478]	[0.655, 0.780]	[33.7, 43.9]	
Gemma-4-E4B	252	0.090	0.122	0.147	0.318	0.422	0.797	36.9	0.0
		[0.021, 0.159]	[0.041, 0.193]	[0.059, 0.226]	[0.239, 0.396]	[0.330, 0.500]	[0.748, 0.843]	[32.0, 41.6]	
Gemma-4-12B	287	0.157	0.198	0.258	0.387	0.477	0.697	34.4	0.0
		[0.089, 0.222]	[0.130, 0.263]	[0.194, 0.327]	[0.326, 0.447]	[0.404, 0.544]	[0.638, 0.754]	[24.9, 41.2]	
Gemma-4-31B	113	0.185	0.085	0.267	0.429	0.401	0.645	42.4	0.6
		[0.080, 0.283]	[−0.051, 0.213]	[0.163, 0.373]	[0.301, 0.543]	[0.273, 0.531]	[0.531, 0.747]	[13.9, 56.1]	
Ministral-3-3B	274	0.020	0.075	0.100	0.182	0.379	0.618	45.5	0.0
		[−0.046, 0.084]	[0.006, 0.135]	[0.020, 0.170]	[0.109, 0.257]	[0.303, 0.448]	[0.557, 0.678]	[38.7, 53.6]	
OLMo-3-7B	282	0.104	0.132	0.134	0.252	0.432	0.606	41.9	0.1
		[0.034, 0.170]	[0.059, 0.193]	[0.067, 0.201]	[0.177, 0.320]	[0.368, 0.490]	[0.535, 0.673]	[33.4, 53.2]	

Table 27: MATH500 and MGSM-zh correct-source profiles, halfway positions (tokens), and recorded 95% intervals.

Model	Task	n	$r(8)$	$r(16)$	$r(32)$	$r(64)$	$m_{1/2}$	u
Qwen3.5-2B	MATH500	90	0.178	0.214	0.309	0.625	48.6	0.8
			[0.094, 0.267]	[0.098, 0.316]	[0.187, 0.414]	[0.528, 0.713]	[40.1, 59.1]	
Qwen3.5-9B	MATH500	63	0.191	0.220	0.295	0.481	> 64	56.4
			[0.050, 0.321]	[0.059, 0.371]	[0.138, 0.437]	[0.311, 0.633]		
Gemma-4-E4B	MATH500	93	0.162	0.225	0.349	0.656	45.0	0.1
			[0.076, 0.251]	[0.127, 0.324]	[0.233, 0.472]	[0.562, 0.751]	[35.0, 53.9]	
Qwen3.5-2B	MGSM-zh	107	0.234	0.324	0.500	0.710	32.0	0.0
			[0.128, 0.336]	[0.219, 0.423]	[0.395, 0.605]	[0.615, 0.795]	[22.0, 42.1]	
Qwen3.5-9B	MGSM-zh	83	0.154	0.331	0.435	0.579	43.7	9.7
			[0.034, 0.271]	[0.175, 0.473]	[0.324, 0.539]	[0.451, 0.693]	[25.6, 61.8]	
OLMo-3-7B	MGSM-zh	119	−0.011	0.125	0.291	0.610	50.4	1.0
			[−0.134, 0.098]	[−0.008, 0.251]	[0.165, 0.410]	[0.518, 0.697]	[41.0, 60.1]	

G Success Variation Along Source Paths

This appendix follows the conditional success probability along individual source paths, whose changes Theorem 1 converts into the coverage cost of a shared interval: six illustrative paths, source-conditioned averages in four models, and the positions at which 2,417 correct paths first reach a high continuation success fraction.

G.1 Continuation measurements

The source-path measurements describe how the probability of a successful continuation changes as a stored solution is revealed. Each prefix receives eight continuations at temperature 1.0, top-$p = 0.95$, and a fresh budget of 320 tokens. Correctness is determined by the final parsed answer, with unparsed answers counted as incorrect. Correct and parsed incorrect source paths are sampled from a question's bank of independent samples; the empirical baseline is its success fraction among

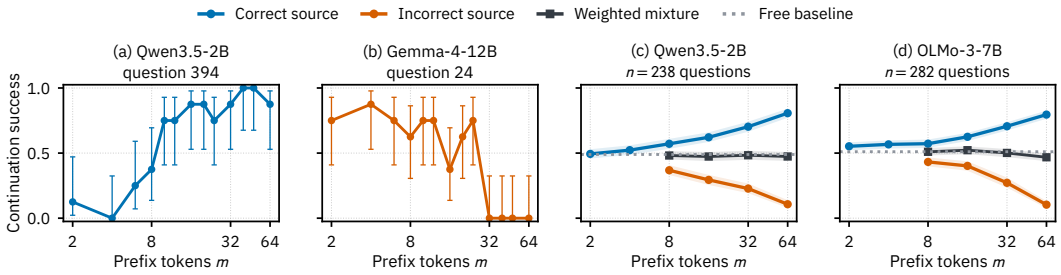

Figure 12: Continuation success along source paths: (a,b) illustrative paths; (c,d) source-conditioned means and weighted mixtures. Bars: Wilson 95% intervals; bands: question-bootstrap 95% intervals.

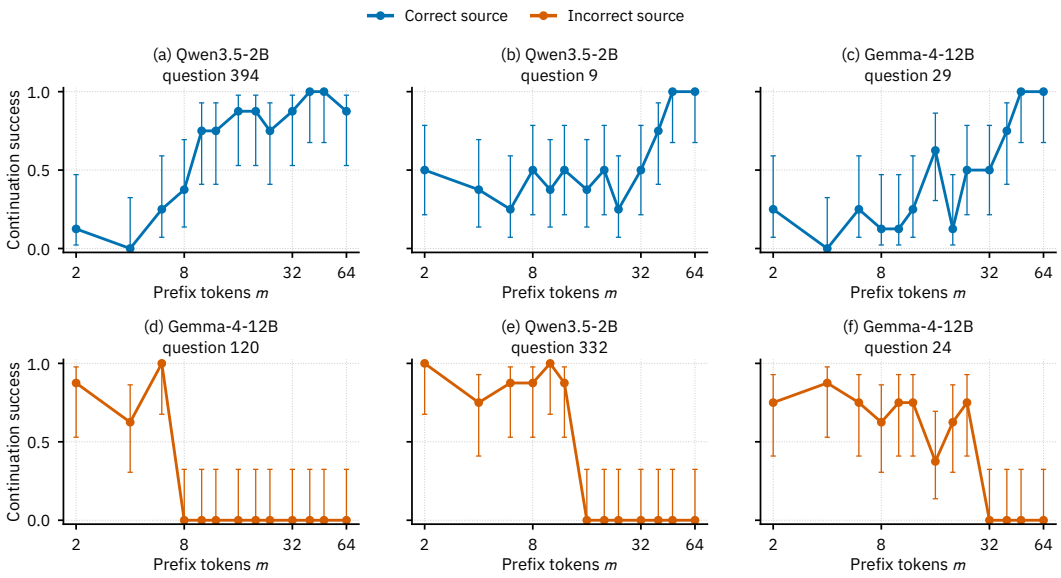

Figure 13: Six illustrative source paths on GSM8K, on a logarithmic token axis: (a)–(c) correct sources; (d)–(f) parsed incorrect sources. Bars are Wilson 95% intervals from eight continuations per prefix.

sixteen independent candidates. These measurements complement the single-prefix measurements of Appendix F.3, which use exact token prefixes and the remaining allowance.

Figure 13 shows all six illustrative paths underlying the examples in Figure 12. Each path is measured at thirteen positions, $m \in \{2, 4, 6, 8, 10, 12, 16, 20, 24, 32, 40, 48, 64\}$, using the same source string throughout and the Hugging Face Transformers implementation. The six paths were chosen after the measurements: within each group of direction and position, the medoid among complete curves with a change of at least 0.45 and a two-level fit of $R^2 \geq 0.65$. Error bars are pointwise Wilson 95% intervals for eight Bernoulli trials.

The paths change in both directions and at different positions. For Qwen3.5-2B on question 394, the measured success fraction rises from $3/8$ at eight tokens to $6/8$ at ten. For Gemma-4-12B on question 24, it falls from $6/8$ at 24 tokens to $0/8$ at 32. These are panels (a) and (b) of Figure 12; the full set also contains later upward changes and earlier downward changes.

G.2 Source-conditioned averages

Figure 14 aggregates one correct and one parsed incorrect source per question in four GSM8K models. Cohorts contain 238, 277, 247, and 282 questions for Qwen3.5-2B, Qwen3.5-9B, Gemma-

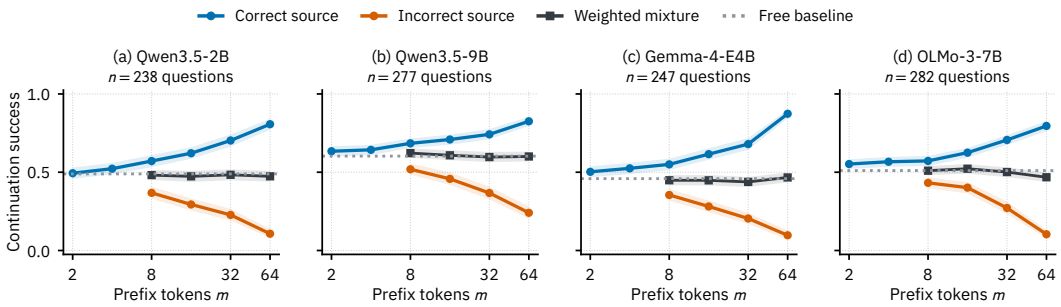

Figure 14: Source-conditioned continuation success and per-question weighted mixtures in four GSM8K cohorts. Bands: 95% question-bootstrap intervals. Dotted lines: cohort mean success of independent samples.

4-E4B, and OLMo-3-7B, respectively. Both source types have eight continuations at each of $m = 8, 16, 32, 64$; correct-source measurements additionally cover $m = 2, 4$. For question q, let $\widehat{\mu}_q$ be the success fraction of the independent bank and $\widehat{a}_C(q, m)$ and $\widehat{a}_W(q, m)$ the two source-conditioned continuation fractions. The plotted mixture averages

$$\widehat{\mu}_q \widehat{a}_C(q, m) + (1 - \widehat{\mu}_q)\widehat{a}_W(q, m)$$

over questions, with the weights applied within each question. Bands are the recorded 95% question-bootstrap intervals from 2,000 replicates, and the dotted line is the cohort's mean success fraction of the independent bank.

The correct-source mean rises and the incorrect-source mean falls in all four models, while their weighted mixture varies less. At 64 tokens, mixture minus baseline is -1.54, -0.23, $+0.76$, and -4.28 pp in model order. The corresponding 95% intervals are $[-3.36, 0.28]$, $[-2.30, 1.78]$, $[-1.07, 2.75]$, and $[-6.24, -2.29]$ pp. The OLMo-3-7B deviation appears in Figure 12, where its weighted mixture lies below the baseline at this position.

Under the matched policy of Section 3, the two source-conditioned means are determined by the variance of $V_m = \mathbb{E}[Z \mid \mathcal{F}_m]$. For $0 < \mu_q < 1$, define $a_C(q, m) = \mathbb{E}[V_m \mid Z = 1, q]$ and $a_W(q, m) = \mathbb{E}[V_m \mid Z = 0, q]$. Since $\mathbb{E}[V_m Z \mid q] = \mathbb{E}[V_m^2 \mid q]$, conditioning gives

$$\begin{aligned} a_C(q, m) &= \mu_q + \frac{\mathrm{Var}(V_m \mid q)}{\mu_q}, \\ a_W(q, m) &= \mu_q - \frac{\mathrm{Var}(V_m \mid q)}{1 - \mu_q}. \end{aligned} \tag{17}$$

Consequently, $\mu_q a_C + (1 - \mu_q)a_W = \mu_q$ and $a_C - a_W = \rho_q(m)$: the gap between the two source-conditioned means is the fraction of success variance that the prefix explains, and the pooled profile in Equation (8) weights these per-question gaps by $\mu_q(1-\mu_q)$. In the measurements above, the fresh continuation budget and the restriction to parsed incorrect sources define the strata, and the mixture residuals quantify their departure from the free baseline.

G.3 Variation across 2,417 correct source paths

Figure 15 shows the complete correct-source matrix for ten GSM8K models. Each question supplies one fixed correct source with eight continuations at all six positions $m \in \{2, 4, 8, 16, 32, 64\}$, 2,417 paths in total, with all six prefixes of a path cut from the same source string. Within each model, rows follow the first position whose measured success fraction is at least $6/8$, then the path's average fraction and question identifier; paths without such a reading by 64 tokens appear below the grey separator. All model panels have equal height, so row height scales with cohort size. Four models were sampled with Hugging Face Transformers at all positions, the other six with vLLM at positions 2 and 4 and Transformers at positions 8 to 64.

The ordering displays the range of positions at which source paths attain a high continuation success fraction. Every model has first readings of at least $6/8$ at all six measured positions, and the model

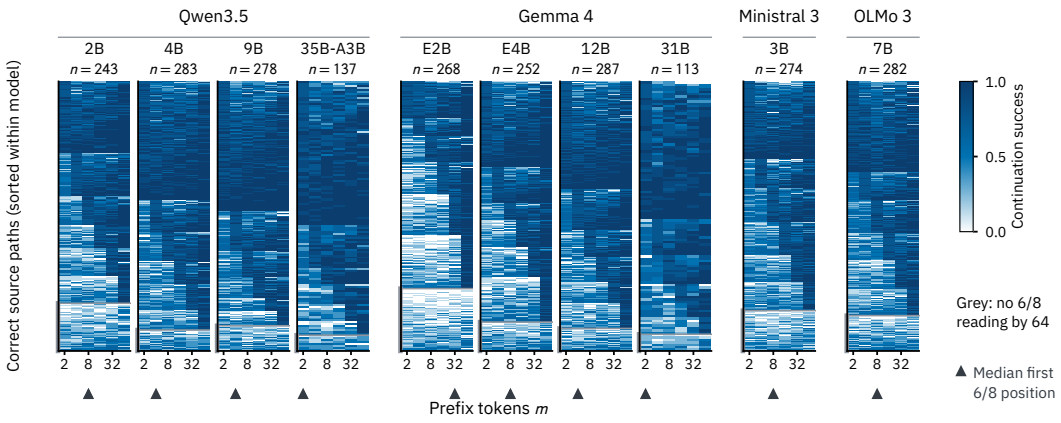

Figure 15: Continuation success on 2,417 correct source paths. Rows are sorted by the first measured $6/8$ success fraction; triangles show its median position, and grey strips mark paths without a crossing by 64 tokens.

medians range from 2 to 32 tokens. These markers summarize threshold attainment on individual correct sources; the profile $r(m)$ of Appendix F.5 pools the same continuation measurements relative to the independent-sample baseline.

H Prefix Evaluation on Matched Source Paths

This appendix asks how much of the information that a prefix carries about eventual success a learned prefix evaluator extracts, by comparing a process reward model with continuation resampling on identical source paths. Under the generation policy π, any predictor g of the outcome from the question and the first m generated tokens satisfies

$$\mathbb{E}[(Z-g)^2] = \mathbb{E}[(Z-V_m)^2] + \mathbb{E}[(V_m-g)^2], \tag{18}$$

so V_m minimizes the expected squared error, and the pooled profile in Equation (8) is the fraction of within-question outcome variance that the prefix explains; its growth locates where information about eventual success becomes available.

We compare continuation-based scores with a process reward model (Lightman et al., 2024) on matched source paths. Each question contributes one correct and one incorrect source path, with measurements complete at 8, 16, 32 and 64 generated tokens. The four GSM8K cohorts contain 238, 277, 247 and 282 questions for Qwen3.5-2B, Qwen3.5-9B, Gemma-4-E4B and OLMo-3-7B, respectively. Continuation resampling scores each prefix by the fraction of eight fresh continuations that succeed. The process reward model, Qwen2.5-Math-PRM-7B, scores the same source prefix at the same token position without recalibration; its training scores reasoning-step boundaries, whereas these positions cut prefixes at fixed token counts. Both scores are evaluated against the final correctness of that source path, and Figure 16 reports pooled empirical AUC on these paired paths.

Across positions, the continuation-based scores exceed the same-path PRM scores by 0.06 to 0.18 AUC, and by 0.09 to 0.18 at 64 tokens. The learned evaluator thus recovers part of the information that continuation resampling reads from the same prefix. Appendix K.1 extends the comparison to eight policy models and to entropy features and hidden-state probes.

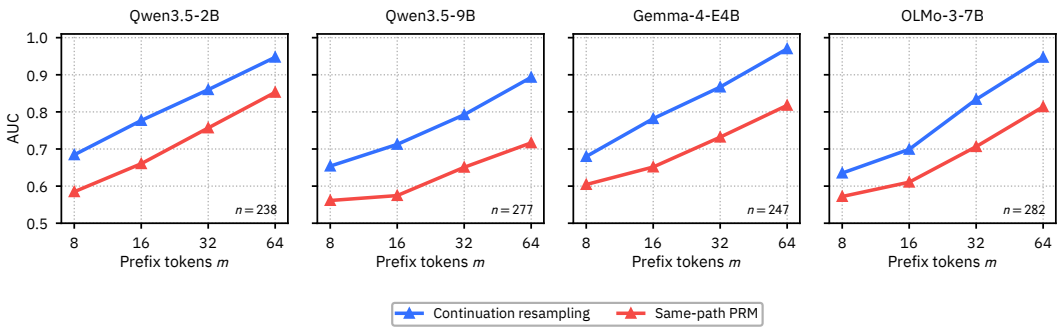

Figure 16: AUC for eventual correctness on identical source paths. Each panel reports one GSM8K model; n counts paired questions.

I Selected Prefixes

This appendix states and proves the decomposition that Section 6 uses for a prefix chosen by a rule instead of sampled from the policy, and gives the setup and all estimates of the greedy-prefix experiment.

I.1 The effect of a selected prefix

Let h_q be a chosen prefix of w tokens with positive probability under π, so that its conditional success probability $V_w(h_q) = \Pr(Z = 1 \mid q, h_q)$ is defined by the policy's continuation distribution. Conditional on the selected prefix and on all information used to select it, the N continuations are fresh independent draws from the continuation policy; selection may use earlier pilot continuations, which are then not reused as the evaluated continuations. The N-candidate coverage of h_q is $1 - \phi_N(V_w(h_q))$.

Corollary 1 (selected prefix)**.** *For any chosen prefix h_q and integer $N \geq 1$,*

$$C_N(0) - \big[1 - \phi_N(V_w(h_q))\big] = \phi'_N(\mu_q)\,(V_w(h_q) - \mu_q) + D_{\phi_N}(V_w(h_q), \mu_q). \tag{19}$$

For a randomized selection rule, the identity holds for each selected prefix and is averaged over the selection randomness. For a prefix drawn from π, averaging over its draw removes the linear term and recovers Equation (5) at $m = 0$, $m' = w$.

For each question, coverage improves exactly when $V_w(h_q) > \mu_q$, since $1 - \phi_N$ is increasing. For $N > 1$ it is strictly concave: improvements on easy questions can add little coverage, while declines elsewhere remove more, so the distribution of changes across questions matters alongside their average.

Proof of Corollary 1. Let h_q be a prefix of w tokens produced by a rule that may depend on q and on the policy but not on the verifier outcome of any candidate continued from it, and let $V_w(h_q) = \Pr(Z = 1 \mid q, h_q)$. Continuing N candidates independently from h_q makes them all fail with probability $\phi_N(V_w(h_q))$, so their coverage at q is $1 - \phi_N(V_w(h_q))$. Since $V_0 = \mu_q$ gives $C_N(0) = 1 - \phi_N(\mu_q)$,

$$\mathbb{E}_q\big[C_N(0)\big] - \mathbb{E}_q\big[1 - \phi_N(V_w(h_q))\big] = \mathbb{E}_q\big[\phi_N(V_w(h_q)) - \phi_N(\mu_q)\big].$$

Expanding the integrand at the base point μ_q by the definition of D_{ϕ_N},

$$\phi_N(V_w(h_q)) - \phi_N(\mu_q) = \phi'_N(\mu_q)\,(V_w(h_q) - \mu_q) + D_{\phi_N}(V_w(h_q), \mu_q),$$

which gives (19) for each question and its corresponding average over questions. The two terms combine to $\phi_N(V_w(h_q)) - \phi_N(\mu_q)$. Since ϕ_N is strictly decreasing for every integer $N \geq 1$, this difference is negative exactly when $V_w(h_q) > \mu_q$. Thus a per-question increase in conditional success probability improves coverage on that question.

Table 28: Greedy prefixes at all five lengths: coverage of 16 candidates and pass@1. Bold: the 95% question-bootstrap interval of the difference from independent sampling excludes zero.

			Greedy prefix, by length w				
Model	Task	Ind.	4	8	16	32	64
Coverage of 16 candidates (%); Ind.: independent pass@16							
Qwen3.5-2B	GSM8K	89.5	88.6	**83.2**	**80.1**	**75.2**	**70.5**
Qwen3.5-4B	GSM8K	99.0	**97.7**	**94.7**	**90.9**	**91.0**	**86.5**
Qwen3.5-9B	GSM8K	100.0	**98.3**	**97.9**	**97.1**	**95.6**	**92.3**
Ministral-3-3B	GSM8K	95.0	95.0	91.9	**89.9**	**87.6**	**80.7**
OLMo-3-7B	GSM8K	94.0	91.9	**90.7**	**89.1**	**84.1**	**74.9**
Qwen3.5-2B	MATH	52.5	**57.1**	**58.8**	55.3	50.0	**41.6**
Qwen3.5-4B	MATH	63.0	62.6	61.5	**57.4**	**55.1**	**49.4**
Qwen3.5-9B	MATH	48.5	51.3	50.5	48.7	46.2	**43.5**
Ministral-3-3B	MATH	61.5	61.7	61.6	60.4	57.2	**52.3**
OLMo-3-7B	MATH	46.0	49.4	48.2	45.5	43.2	**37.5**
pass@1*: independent (%), and greedy-prefix minus independent (pp)*							
Qwen3.5-2B	GSM8K	40.9	**+3.0**	**+3.9**	**+5.6**	**+8.2**	**+16.0**
Qwen3.5-4B	GSM8K	69.4	**+2.5**	**+2.5**	**+4.1**	**+5.5**	**+7.5**
Qwen3.5-9B	GSM8K	83.1	**+1.6**	**+2.2**	**+3.1**	**+3.0**	+2.2
Ministral-3-3B	GSM8K	50.2	**+4.6**	**+6.5**	**+8.5**	**+11.5**	**+14.1**
OLMo-3-7B	GSM8K	53.2	+2.0	**+3.3**	**+5.1**	**+7.9**	**+10.0**
Qwen3.5-2B	MATH	14.9	**+2.0**	**+3.1**	**+5.2**	**+6.2**	**+9.8**
Qwen3.5-4B	MATH	20.2	+1.0	+1.3	+1.3	**+2.2**	**+4.8**
Qwen3.5-9B	MATH	19.2	−0.2	−0.6	−0.4	−0.2	+0.4
Ministral-3-3B	MATH	18.8	+1.2	+1.4	**+2.4**	**+5.5**	**+9.8**
OLMo-3-7B	MATH	11.6	+0.5	+0.6	**+1.5**	**+3.2**	**+5.6**

Across questions, the concavity of $f_N(v) = 1 - (1 - v)^N$ makes the locations of improvements and declines consequential. For example, with $N = 2$, two equally weighted questions with independent success probabilities $(0.2, 0.8)$ have average individual success 0.5 and average coverage 0.66. Chosen prefixes with probabilities $(0.1, 1.0)$ increase average individual success to 0.55 and reduce average coverage to 0.595.

If instead the prefix is drawn from the policy, $h_q \sim \pi(\cdot \mid q)$ restricted to its first w tokens, then $V_w(h_q) = V_w$ and Proposition 4 gives $\mathbb{E}[V_w \mid q] = \mu_q$, so the first term vanishes question by question and the second is the right-hand side of (5) at $m = 0$ and $m' = w$, averaged over questions. □

I.2 Greedy prefixes

Setup. A greedy prefix consists of the first w tokens decoded at zero temperature. We draw 32 sampled continuations from each prefix and estimate its coverage at $N = 16$ with Equation (12). The experiment covers five models on GSM8K and the numeric MATH subset at $w \in \{4, 8, 16, 32, 64\}$, giving 50 model–task–length combinations. The cost is independent pass@16, sampled at temperature 1 throughout generation, minus estimated greedy-prefix coverage; positive values are losses. Table 28 reports all fifty estimates with the corresponding changes in pass@1; intervals are 95% bootstrap intervals over the 200 questions of each setting.

Results. Section 6 and Figure 6 summarize these estimates. At 64 tokens, the share of questions on which none of the 32 continuations succeeds, although at least one of the 16 independent samples does, is 6.5 to 21.0% on GSM8K and 8.0 to 16.5% on MATH. Appendix J reports prefixes sampled at a lower temperature.

Table 29: Prefix-temperature length scan: pass@1 / pass@16 (%). Prefix temperature is 0.3, continuation temperature 1.0.

Model	Task	Baseline	$w = 4$	$w = 8$	$w = 16$	$w = 32$	$w = 64$
Qwen3.5-2B	GSM8K	40.9 / 89.5	42.5 / 84.0	44.9 / 88.0	46.3 / 88.5	48.5 / 85.0	52.3 / 85.0
Qwen3.5-4B	GSM8K	69.4 / 99.0	71.1 / 98.0	72.9 / 95.0	72.8 / 97.0	75.1 / 97.0	76.8 / 95.0
Qwen3.5-9B	GSM8K	83.1 / 100.0	83.9 / 99.0	84.7 / 99.0	86.3 / 98.0	85.9 / 98.0	87.0 / 97.5
Ministral-3-3B	GSM8K	50.2 / 95.0	53.6 / 92.5	56.2 / 93.5	59.8 / 92.5	59.8 / 96.0	64.6 / 94.0
OLMo-3-7B	GSM8K	53.2 / 94.0	56.5 / 95.0	57.7 / 94.0	58.4 / 94.5	62.2 / 93.5	64.7 / 92.5
Qwen3.5-2B	MATH	14.9 / 52.5	16.7 / 57.5	18.3 / 62.0	19.2 / 62.0	21.2 / 62.5	25.8 / 63.5
Qwen3.5-4B	MATH	20.2 / 63.0	20.0 / 57.5	21.7 / 62.5	21.1 / 60.0	23.6 / 61.5	25.8 / 60.5
Qwen3.5-9B	MATH	19.2 / 48.5	19.2 / 51.0	19.2 / 50.5	18.8 / 53.5	20.0 / 51.0	19.8 / 45.5
Ministral-3-3B	MATH	18.8 / 61.5	20.8 / 62.0	20.4 / 63.0	22.2 / 63.0	22.5 / 60.5	26.3 / 61.0
OLMo-3-7B	MATH	11.6 / 46.0	12.0 / 48.0	11.8 / 44.5	14.3 / 51.0	13.5 / 44.5	16.5 / 46.5

J Temperature Interventions

This appendix measures the other operation on a prefix, selection through temperature: every candidate draws its own prefix, so the candidate set keeps 16 prefix draws, and what changes is the distribution they come from. The experiments vary prefix temperature, prefix length, and the position of a fixed higher-temperature region.

J.1 Changing the prefix distribution

Temperature changes the distribution from which a prefix is selected. Write $P(\cdot \mid q)$ for the distribution of a prefix of w tokens, with the continuation policy fixed, and let $\mu_q^P = \mathbb{E}_{h\sim P(\cdot|q)}[V_w(h)]$. When every candidate draws its own prefix, coverage is

$$C_N^{\mathrm{ind},P} = \mathbb{E}_q\big[1 - (1 - \mu_q^P)^N\big]. \tag{20}$$

When all candidates inherit one prefix draw, it is

$$C_N^{\mathrm{shared},P} = \mathbb{E}_{q,\, h\sim P(\cdot|q)}\big[1 - (1 - V_w(h))^N\big]. \tag{21}$$

Thus prefix selection changes the distribution of conditional success probabilities, while sharing changes how many draws from that distribution the candidate set receives. At zero prefix temperature, P is concentrated on the greedy prefix and the two expressions coincide with Corollary 1.

The experiments use 200 questions per setting on GSM8K and the numeric MATH subset, with four-shot prompts, top-$p = 0.95$, and total generation budgets of 320 and 640 tokens, respectively. The prefix-length scan, the four-token comparison, and the runs with a displaced higher-temperature region use the same questions, prompts, and answer evaluation as their constant-temperature baselines and the opening-region run. Intervals use 4,000 paired question-bootstrap replicates.

J.2 Prefix length at temperature 0.3

Each of 16 candidates samples its first $w \in \{4, 8, 16, 32, 64\}$ tokens at temperature 0.3 and then continues at temperature 1.0, with a separate continuation seed for each candidate. The baseline uses temperature 1.0 throughout. Table 29 reports candidate quality and coverage together. Extending the prefix sampled at temperature 0.3 from four to 64 tokens raises pass@1 in all ten settings, by 0.6 to 11.0 pp, while the corresponding pass@16 changes range from -5.5 to $+6.0$ pp. For Qwen3.5-9B on MATH, pass@1 rises from 19.2% to 19.8% and coverage falls from 51.0% to 45.5%; for Qwen3.5-2B on MATH, the same extension raises coverage from 57.5% to 63.5%. The prefix distribution therefore changes candidate quality and collective coverage in task-dependent ways.

J.3 Four-token prefixes and the greedy endpoint

Table 30: Four-token prefixes: candidate quality, coverage, and paired difference ($T = 0.3$ minus greedy). Distinct and modal summarize the 16 sampled prefixes.

Model	Task	$T = 0.3$ @1 / @16 (%)	Greedy @1 / @16 (%)	Δ@16 (pp) [95% CI]	Distinct	Modal (%)
Qwen3.5-2B	GSM8K	42.9 / 86.5	43.9 / 87.7	-1.2 [-5.1, +2.7]	3.20	67
Qwen3.5-4B	GSM8K	71.9 / 98.5	72.0 / 97.0	$+1.5$ [-0.3, +3.4]	2.78	72
Qwen3.5-9B	GSM8K	84.7 / 99.0	84.7 / 98.3	$+0.7$ [-0.5, +2.2]	2.52	76
Ministral-3-3B	GSM8K	54.5 / 95.0	54.5 / 95.8	-0.8 [-3.1, +1.4]	3.04	72
OLMo-3-7B	GSM8K	55.0 / 94.5	54.8 / 92.0	$+2.5$ [+0.6, +4.7]	3.32	67
Qwen3.5-2B	MATH	17.6 / 62.5	17.0 / 57.6	$+4.9$ [-0.3, +10.3]	4.59	58
Qwen3.5-4B	MATH	21.1 / 60.5	20.8 / 61.6	-1.1 [-5.2, +3.0]	3.85	62
Qwen3.5-9B	MATH	18.8 / 51.0	18.8 / 51.8	-0.8 [-5.0, +3.2]	3.08	68
Ministral-3-3B	MATH	20.4 / 65.5	20.1 / 61.0	$+4.5$ [-0.2, +9.1]	4.50	60
OLMo-3-7B	MATH	13.2 / 53.0	12.2 / 48.9	$+4.1$ [-0.4, +8.7]	4.09	61

A separate four-token comparison records the sampled prefix strings. It uses 16 candidates with prefix temperature 0.3, and 32 continuations of the greedy prefix to estimate pass@16 with the combination estimator in (12). Table 30 reports all ten settings, including their paired coverage differences. The finite-temperature arm contains 2.5 to 4.6 distinct prefixes among 16 candidates, and its most frequent prefix accounts for 58 to 76% of the candidates.

Figure 17 compares the two prefix distributions at the same candidate count. The coverage difference, temperature 0.3 minus greedy, ranges from -1.2 to $+4.9$ pp. The largest difference occurs for Qwen3.5-2B on MATH, with a 95% interval of $[-0.3, +10.3]$. OLMo-3-7B on GSM8K gains 2.5 pp with an interval of $[+0.6, +4.7]$; the other nine intervals include zero. Both interventions narrow the prefix distribution, while finite temperature retains several possible prefixes.

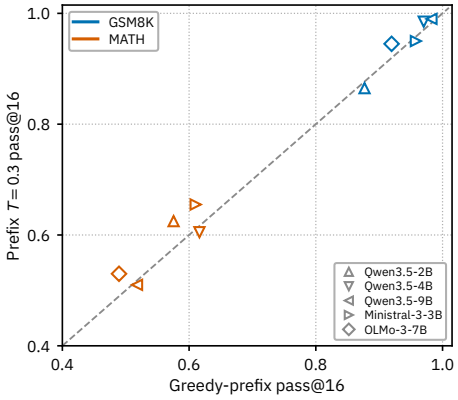

Figure 17: Four-token prefix coverage at temperature 0.3 versus the greedy-prefix estimate; $N = 16$.

Figure 18 compares the changes in candidate quality and coverage against constant temperature 1.0. Candidate quality increases by 0.9 to 4.3 pp in nine settings and decreases by 0.3 pp for Qwen3.5-9B on MATH, while coverage changes range from -3.0 to $+10.0$ pp. Qwen3.5-2B on MATH gains 10.0 pp with an interval of $[+3.5, +16.0]$, and OLMo-3-7B on MATH gains 7.0 pp with an interval of $[+0.5, +13.5]$. The largest decrease is Qwen3.5-2B on GSM8K, -3.0 pp with an interval of $[-8.0, +2.0]$. Table 31 reports both changes and their paired intervals for every setting. Candidate quality and coverage average μ_q^P and $1 - (1 - \mu_q^P)^{16}$, respectively, which makes the distribution of changes across questions relevant to the set's coverage.

J.4 Position of a higher-temperature region

Each of 16 independently generated candidates uses temperature 1.0 for a consecutive region of 16 tokens and temperature 0.3 elsewhere. The region starts after 0, 16, 48, or 96 tokens, denoted by W_0, W_{16}, W_{48}, and W_{96}. Table 32 pairs the opening region W_0 with each displaced region on the same 200 questions.

The coverage difference $W_0 - W_{96}$ ranges from -1.0 to $+5.5$ pp. Its interval excludes zero for Qwen3.5-2B and Qwen3.5-4B on GSM8K, with differences of $+5.5$ and $+3.0$ pp, respectively. For the nearer regions, the $W_0 - W_{16}$ interval excludes zero in one setting and $W_0 - W_{48}$ in two. These interventions move where each candidate's randomness enters, and their position dependence differs from that of the coverage cost of sharing one prefix.

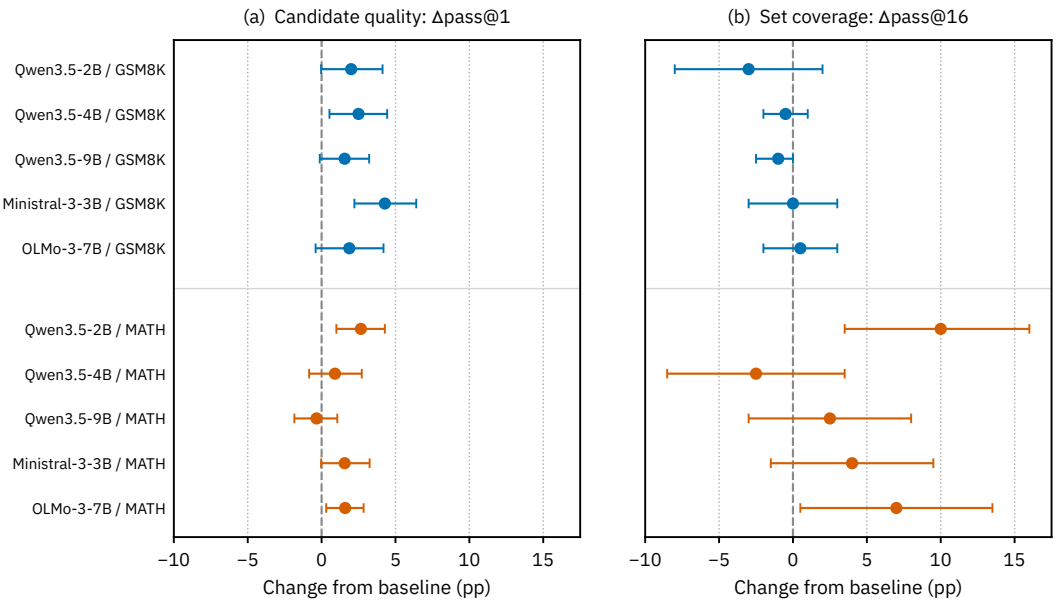

Figure 18: Four-token prefixes at temperature 0.3 relative to constant temperature 1.0: changes in (a) candidate quality and (b) coverage. Bars: 95% paired question-bootstrap intervals.

Table 31: Four-token prefixes at temperature 0.3 versus constant temperature 1.0. Differences use paired question-bootstrap intervals.

Model	Task	Baseline @1 / @16 (%)	Four-token $T = 0.3$ @1 / @16 (%)	Δ@1 (pp) [95% CI]	Δ@16 (pp) [95% CI]
Qwen3.5-2B	GSM8K	40.9 / 89.5	42.9 / 86.5	+2.00 [−0.03, +4.12]	−3.0 [−8.0, +2.0]
Qwen3.5-4B	GSM8K	69.4 / 99.0	71.9 / 98.5	+2.50 [+0.53, +4.44]	−0.5 [−2.0, +1.0]
Qwen3.5-9B	GSM8K	83.1 / 100.0	84.7 / 99.0	+1.56 [−0.12, +3.22]	−1.0 [−2.5, +0.0]
Ministral-3-3B	GSM8K	50.2 / 95.0	54.5 / 95.0	+4.28 [+2.22, +6.41]	+0.0 [−3.0, +3.0]
OLMo-3-7B	GSM8K	53.2 / 94.0	55.0 / 94.5	+1.88 [−0.41, +4.19]	+0.5 [−2.0, +3.0]
Qwen3.5-2B	MATH	14.9 / 52.5	17.6 / 62.5	+2.66 [+1.00, +4.28]	+10.0 [+3.5, +16.0]
Qwen3.5-4B	MATH	20.2 / 63.0	21.1 / 60.5	+0.91 [−0.84, +2.72]	−2.5 [−8.5, +3.5]
Qwen3.5-9B	MATH	19.2 / 48.5	18.8 / 51.0	−0.34 [−1.84, +1.06]	+2.5 [−3.0, +8.0]
Ministral-3-3B	MATH	18.8 / 61.5	20.4 / 65.5	+1.56 [−0.03, +3.25]	+4.0 [−1.5, +9.5]
OLMo-3-7B	MATH	11.6 / 46.0	13.2 / 53.0	+1.59 [+0.31, +2.84]	+7.0 [+0.5, +13.5]

Table 32: Position of a 16-token region at temperature 1.0 with temperature 0.3 elsewhere: pass@16 (%) and paired differences (pp).

Model	Task	W_0	W_{16}	W_{48}	W_{96}	$W_0 - W_{16}$ [95% CI]	$W_0 - W_{48}$ [95% CI]	$W_0 - W_{96}$ [95% CI]
Qwen3.5-2B	GSM8K	89.5	88.0	87.5	84.0	$+1.5$ [-2.5, +6.0]	$+2.0$ [-3.0, +7.0]	$+5.5$ [+0.5, +10.5]
Qwen3.5-4B	GSM8K	98.0	95.5	94.5	95.0	$+2.5$ [+0.5, +5.0]	$+3.5$ [+1.0, +6.0]	$+3.0$ [+1.0, +5.5]
Qwen3.5-9B	GSM8K	98.5	97.5	97.5	98.5	$+1.0$ [-1.0, +3.0]	$+1.0$ [+0.0, +2.5]	$+0.0$ [-1.5, +1.5]
Ministral-3-3B	GSM8K	96.0	93.5	94.0	97.0	$+2.5$ [-1.0, +6.0]	$+2.0$ [-0.5, +5.0]	-1.0 [-3.5, +1.5]
OLMo-3-7B	GSM8K	96.0	94.0	93.5	94.5	$+2.0$ [-0.5, +5.0]	$+2.5$ [-0.5, +6.0]	$+1.5$ [-1.0, +4.0]
Qwen3.5-2B	MATH	68.5	64.5	64.0	63.5	$+4.0$ [-1.5, +9.5]	$+4.5$ [-1.5, +10.5]	$+5.0$ [-1.0, +11.0]
Qwen3.5-4B	MATH	58.5	61.0	54.0	55.0	-2.5 [-8.0, +2.5]	$+4.5$ [-0.5, +9.5]	$+3.5$ [-1.5, +9.0]
Qwen3.5-9B	MATH	46.0	43.0	39.0	43.5	$+3.0$ [-2.0, +8.0]	$+7.0$ [+3.0, +11.5]	$+2.5$ [-2.0, +7.5]
Ministral-3-3B	MATH	68.0	67.0	66.5	68.0	$+1.0$ [-4.0, +6.5]	$+1.5$ [-4.0, +7.0]	$+0.0$ [-6.0, +6.0]
OLMo-3-7B	MATH	43.0	43.5	41.5	41.5	-0.5 [-6.5, +5.5]	$+1.5$ [-4.5, +7.5]	$+1.5$ [-3.5, +6.5]

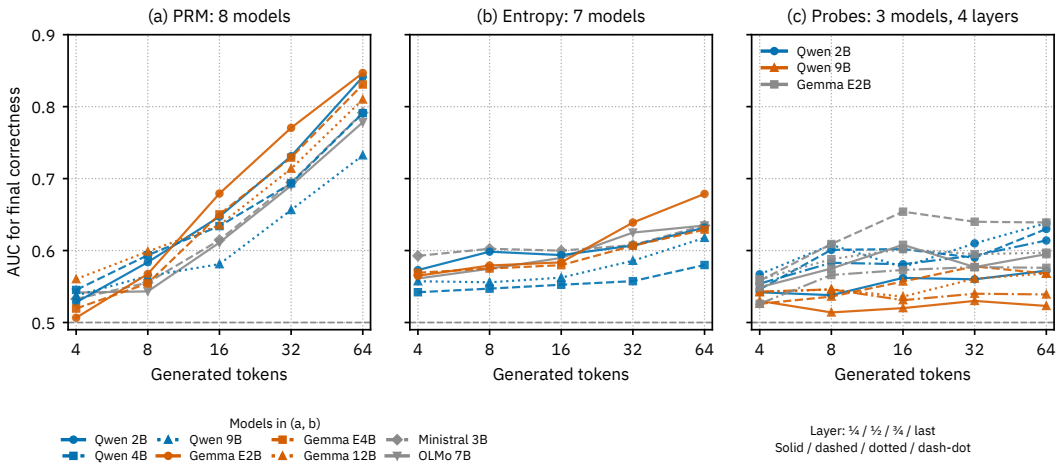

Figure 19: AUC for eventual correctness versus prefix length: PRM, entropy features, and linear probes on their respective GSM8K cohorts. Each curve is a model or model–layer series.

K Additional Experiments

This appendix adds a measurement outside our sampling design: how three prefix signals rank eventual correctness across more models, which extends Appendix H.

K.1 Prefix signals across broader model populations

The matched comparison in Appendix H evaluates two scores on identical source paths. We extend it here to more models with recorded prefix evaluations on independently sampled GSM8K candidates. All three signal families predict the final correctness of the same candidate at 4, 8, 16, 32, and 64 generated tokens. The reported AUC is the mean over five test folds grouped by question.

The uncalibrated Qwen2.5-Math-PRM-7B scores eight policy models: Qwen3.5-2B/4B/9B, Gemma-4-E2B/E4B/12B, Ministral-3-3B, and OLMo-3-7B. Its score is the positive-class probability at a separator appended to the token prefix. The entropy-feature evaluation covers the same models except Gemma-4-12B and uses token entropy at the current position together with its prefix mean, maximum, and standard deviation. Linear hidden-state probes cover Qwen3.5-2B, Qwen3.5-9B, and Gemma-4-E2B at one-quarter, one-half, three-quarter, and full model depth, producing twelve model–layer series.

The PRM and probe evaluations retain candidates with at least 64 tokens. The entropy evaluation first truncates at a subsequent question marker and requires at least 66 tokens. This yields 3,969–4,800 candidates per PRM model (36,039 total) and 3,231–4,571 per entropy model. Figure 19 therefore reports each family's position dependence on its own cohort; the matched comparison remains in Appendix H.

The PRM's AUC rises from 0.507–0.560 at four tokens to 0.732–0.847 at 64 tokens across the eight models. Entropy-feature AUCs span 0.542–0.679 over all measured positions, and the outcome probes span 0.514–0.654. The conditional-success profile characterizes the information available in the prefix, and these curves measure how much of that information each signal extracts at each prefix length.

L Related Work

Multi-candidate reasoning and coverage. Repeated sampling raises coverage in code generation and across reasoning tasks (Chen et al., 2021; Brown et al., 2024), majority voting over sampled paths raises accuracy without a verifier (Wang et al., 2023), and inference-time scaling studies allocate generation among sampling, search and revision (Snell et al., 2025; Wu et al., 2025). Coverage

at large candidate counts also serves to compare training methods (Yue et al., 2025; Hayes et al., 2026). These studies vary how many candidates are generated. We vary how much of their generation the candidates share, under a fixed policy and an equal budget.

Shared generation in search and training. Tree search over partial solutions (Yao et al., 2023; Ding et al., 2025a; Li et al., 2026b), forking at branch points chosen by model confidence or learned (Al-Khalili et al., 2026; Jia et al., 2026), and tree-structured rollouts for reinforcement learning (Li et al., 2026a; Hou et al., 2025; Li et al., 2025) reuse prefixes to save generation, training and serving systems make the reuse cheap (Yu et al., 2026a; Zheng et al., 2024; Juravsky et al., 2024), learned parallel reasoning lets the model decide when to spawn and merge threads (Pan et al., 2025; Yang et al., 2025), and collaborative parallel thinking instead broadcasts intermediate results across independent branches or lets concurrent workers attend to one another (Wang et al., 2026; Rodionov et al., 2025). These methods combine sharing with selection, pruning or training and are evaluated end to end; we isolate the effect of sharing a sampled prefix under a matched policy and an equal budget. With long reasoning on competition problems, sharing combined with a choice of branch points has paid off. TreeRL's trees, forked at the most uncertain tokens of generated responses over two iterations, reach a higher PassRate, the coverage of all their responses, than independent sampling with a similar number of generated tokens, and a higher one than trees forked at random (Hou et al., 2025); Fork-Think forks a greedy seed path at its least confident tokens and matches or exceeds the majority-vote accuracy of independent parallel thinking with up to 30% fewer tokens (Al-Khalili et al., 2026). Branch points chosen with the text generated after them, greedy prefixes and trees of several levels lie outside Theorem 2, Section 6 measures how a chosen prefix changes the success probability that its candidates inherit, and our matched measurements do not reach this regime. Path-consistency and prefix-consensus methods continue the prefixes that most sampled paths share (Zhu et al., 2024; Jindal et al., 2026), a selection whose gain in inherited success Corollary 1 weighs against the sharing cost, and concurrent work adapts tree-structured rollouts to question difficulty (Yu et al., 2026b). Outside language models, multilevel splitting and RESTART in rare-event simulation weigh the correlation that shared trajectory segments induce against the cost they save, and branch only where a trajectory has made progress (Villén-Altamirano & Villén-Altamirano, 1991; Glasserman et al., 1999); restart schedules for randomized algorithms (Luby et al., 1993) are the counterpart of Proposition 3.

Where outcomes are decided. Resampling at individual tokens reveals forking tokens after which outcomes change (Bigelow et al., 2025); alternatives at the first decoded token already contain distinct reasoning paths (Wang & Zhou, 2024); resampling the sentences of a trace identifies the steps that most change its answer (Bogdan et al., 2025); return variance in reasoning concentrates at critical forks (Li, 2026); answers often stabilize at a commitment boundary well before a trace ends (Scalena et al., 2026); reinforcement learning with verifiable rewards narrows the solution space mainly at the entrance of a trajectory (Zhou & Li, 2026); and high-entropy tokens steer reasoning directions (Wang et al., 2025). These works locate decisive tokens within individual trajectories or training updates; we measure the success variance that a prefix resolves across sampled trajectories and convert it into the coverage that a shared set loses.

Conditional success and prefix evaluation. Process supervision and rollout-based labels estimate the success probability of partial solutions to train evaluators, guide search or assign credit in reinforcement learning (Lightman et al., 2024; Wang et al., 2024; Luo et al., 2024; Kazemnejad et al., 2025; Zhou et al., 2026). These estimates target the mean of V_m for a given prefix. The coverage of shared candidates depends on its variance across sampled prefixes, which continuation resampling estimates from the same rollouts.