WHAT DO EARLY ANSWERS REVEAL? READOUT PROTOCOLS AND POST-ANSWER REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Does a correct continuation from an early reasoning prefix mean the model was ready to stop? We compare free continuation (no extra request) with continuation prompted to give a final answer, scoring both first and last answers. Across Qwen3 and GPT-OSS on MATH-500 and GPQA-Diamond, recovery depends on the request and which answer is scored. On 100 MATH problems, Qwen3-8B with thinking enabled at a 10% prefix achieves 69.3% free last-answer accuracy; requested outputs rise from 31.8% at the first answer to 73.8% at the last. Among improvements, the median delay from the first answer to the first later correct box is 1,258 tokens. Allowing 16,384 reference tokens on 50 reused questions preserves a 52-point first-to-last gain, but no clear last-answer advantage for either procedure. Closing the reasoning channel reduces later gains in several settings, but not clearly for Qwen at 50%. At 50%, equal continuation budgets retain a Qwen prefix benefit with thinking, but not a clear benefit without it. An online case study saves tokens in the main generation while increasing time and total generation. Recovery, prefix benefit, and stopping efficiency therefore require separate evidence: score first and last answers, compare question-only recovery, and count the full cost before interpreting recovery as readiness to stop.

1 INTRODUCTION

Giving a model more computation at test time can improve reasoning accuracy (Snell et al., 2024), but models may continue after producing an answer. Early exit aims to stop the written chain of thought (CoT) once enough information has been generated. A black-box test continues a short initial segment (*prefix*) and inspects the answer (Wang et al., 2025; Mo et al., 2026). Does a correct continuation show that the model was ready to stop?

A model might first give a wrong answer, continue reasoning, and then correct it. Scoring only that final answer would hide the extra work. Short answer-request continuations (probes) can agree before the original trajectory corrects itself (Mo et al., 2026), and continuation budgets can alter apparent progress (Bulut, 2026). We call a procedure that continues a prefix and selects an answer a *readout*. Early-answer recoverability is a property of this protocol, not direct evidence that the prefix contains a ready answer or that stopping saves time. Figure 1 summarizes the separation.

We compare free continuation, which adds no request, with answer-forced continuation, which appends a final-answer request (suffix). Matching sampling and output budget (allowed new tokens), we score both first and last extracted answers. This separates the request from the scored endpoint using generated text.

The audit separates three claims that a successful continuation can otherwise mix together: the first requested answer is correct, the prefix helps obtain it, and probing makes stopping worthwhile. For Qwen3-8B with thinking enabled (Think), requested outputs gain 42.0 accuracy points after their first answer. Explicitly switching from reasoning to the final-response channel removes much of this gain at 10%, but not clearly at 50%. Longer original traces (references) preserve revision without preserving the original free-versus-forced ranking. The robust finding is the contribution of later generation; the ranking depends on the protocol.

Prefix benefit and stopping efficiency need their own tests. At the 50% checkpoint, equal continuation budgets leave no clear prefix advantage with thinking disabled (NoThink), while Think retains a positive advantage. In an online NoThink case study, agreement-based stopping saves 31% of main trace tokens but takes 2.59× the wall time once probes are counted. These three audits clarify what readouts establish and what remains to be tested for an efficient stopping policy.

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

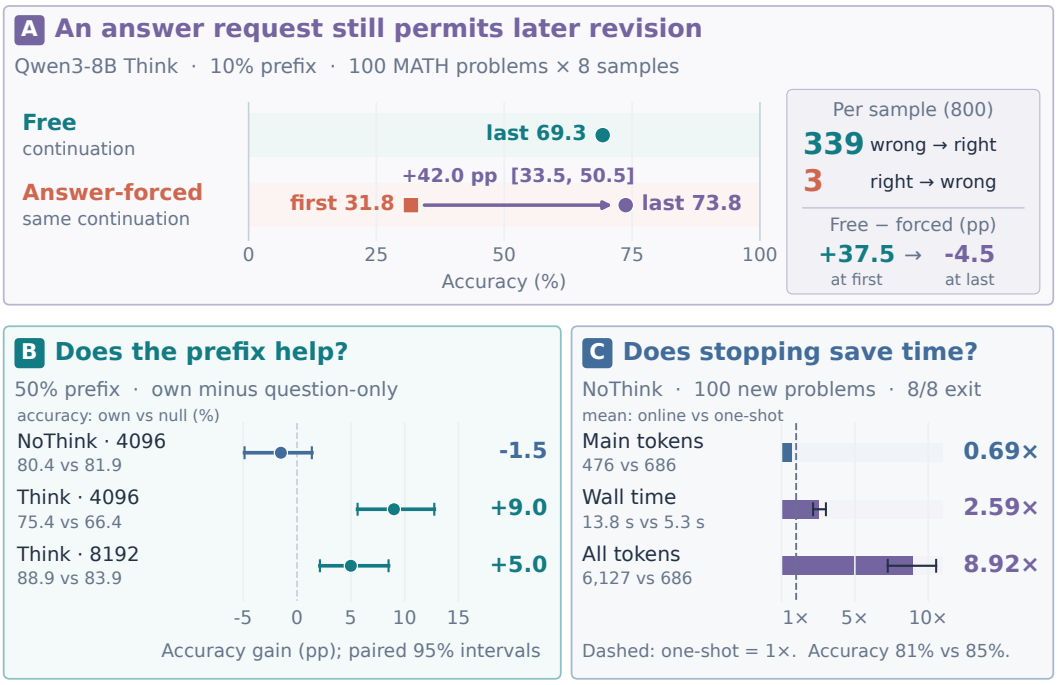

Figure 1: **Separate answer recovery, prefix benefit, and online cost.** Qwen3-8B (8B means 8 billion parameters) on the mathematics benchmark MATH-500; Think/NoThink enable/disable thinking. (A) Score first/last answers under matched settings; inset: per-sample changes. The reversal is concentrated in references that reach their token limit. (B) Own-prefix minus question-only (null) accuracy at 50% (gray: both accuracies), for both modes and common token limits; intervals resample matched problems together. Think retains a positive advantage; some question-only outputs still hit the cap. (C) Exit on eight identical, nonempty continuation answers (probes). Accuracy is per sample; pp means percentage points; one-shot means uninterrupted generation. Costs are ratios of means versus one-shot (whiskers: 95% intervals; gray: means). Wall time and all tokens include probes. Offline fractions use observed, possibly truncated reference lengths.

1. **Tracking answer changes.** Across 14,400 main-study continuations and 9,984 pre-specified follow-up continuations, we separate first answers, later answer changes, missing labels, and revision time. This shows exactly what final-answer accuracy hides.
2. **Testing evaluation choices.** Controls that match the scored answer, token budget, stopping condition, question, and grading identify when a readout changes the conclusion. Request, channel-closure, longer-reference, and question-only controls test whether revision persists and whether the prefix improves recovery.
3. **A deployment case study.** We evaluate one observable online agreement rule with fixed limits on parallel requests, reporting accuracy, wall time, tokens in the main generation, and total generation including probes.

2 Related Work

Closest studies. Two recent studies ask related questions. Bulut (2026) compares own-prefix continuations with matched-budget restarts, asking whether the prefix saves work or carries answer-specific information. Mo et al. (2026) tests whether repeated answer agreement reliably predicts when reasoning can stop; our online audit supports the same practical caution. Our main audit instead tracks where a continuation's final score comes from: its first answer, a newly readable label, or later revision. We compare matched free and requested outputs, including missing answers and revision times.

Probing reasoning progress. Probes of internal states predict answers or correctness before a CoT ends (Boppana et al., 2026; Zhang et al., 2025; Cox et al., 2026); Chain-of-Probe requests intermediate answers and tracks their token probabilities (Wang et al., 2025). Thought Anchors uses

forced answers and continuation resampling to study sentence-level importance (Bogdan et al., 2025). These methods ask what a prefix predicts or which step matters. We instead audit the generated readout itself, including what happens after its first answer and how the scored endpoint, output budget, and grader change the result.

Budget forcing and probe-based stopping. Controlling the reasoning boundary is established practice: s1 lengthens thinking with "Wait" or ends it through budget forcing (Muennighoff et al., 2025); NoThinking bypasses explicit thinking and studies accuracy under token and time limits (Ma et al., 2025). TALE puts token budgets in the prompt to reduce reasoning length (Han et al., 2025). DEER requests trial answers, evaluates their token confidence, and then ends thinking when its threshold is met (Yang et al., 2025b). Dynasor-CoT probes across checkpoints, tests answer consistency, and checks post-answer uncertainty words (Fu et al., 2024; 2025). These methods already control reasoning boundaries. We measure the first-to-last answer change that remains after closing reasoning, across models and checkpoints. Appendix J compares their protocols; we do not assume they use our long free probes or score the last answer.

Early answering and continuation agreement. Lanham et al. (2023) shorten CoT and request early answers to test faithfulness. Successful recovery alone need not identify what the prefix contributed, motivating our question-only controls. Self-consistency combines sampled paths (Wang et al., 2022); Prefix Consistency samples new continuations to weight votes (Iwase et al., 2026). They combine multiple answers; we track changes within one output. Additional voting experiments vary how many answers share a starting prefix; these checks are reported separately (Appendix N). Our online case study illustrates complete cost accounting for one rule, not a performance comparison with DEER or Dynasor-CoT.

Revision after an answer. Self-Refine uses separate feedback and revision steps (Madaan et al., 2023), while self-correction without external feedback can fail on reasoning tasks (Huang et al., 2024). We measure answer changes within one continuation, without an extra critique request.

3 Experimental Design

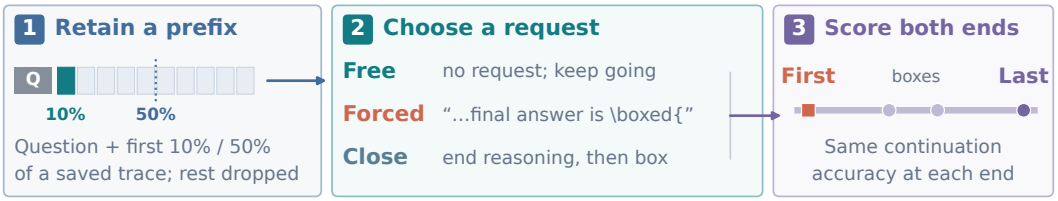

Figure 2: **Readout procedure.** Retain a prefix of a saved reference trace, vary the request, and score the first and last answers in the same continuation. Free adds no request; Forced requests an answer; Close ends reasoning first (follow-ups only). Fractions use observed trace lengths.

Figure 2 separates continuation from scoring. We ask three questions: Does the continuation revise its first answer? Does the prefix help beyond the question alone? Does probing save time when used to stop generation? We first generate a reference trace, retain an early prefix, and sample continuations from that prefix. Gold answers (benchmark ground truth) score outputs but never select reference traces or determine online stopping. The application programming interface (API) must support continuing a supplied assistant prefix; not every chat endpoint does.

3.1 Reference Traces and Prefixes

For each problem indexed by j, let T_{obs} be the observed reference length in tokens and $y = (y_1, \ldots, y_{T_{\text{obs}}})$ its token sequence, with y_t denoting token t. We sample at temperature 1; temperature controls sampling randomness. The trace ends at the model's end-of-sequence (EOS) marker or the reference cap. We call the latter case *capped* (censored in tables): its natural completion length is unknown. We retain either 10% or 50% of the observed trace; for example, a 2,000-token trace gives prefixes of 200 and 1,000 tokens. Formally, for retained fraction $f \in \{0.10, 0.50\}$, the offline prefix $c^{(f)} = y_{1:k_f}$ contains the first k_f tokens, where $k_f = \max(1, \min(\lfloor f T_{\text{obs}} \rfloor, T_{\text{obs}} - 1))$. Here $\lfloor x \rfloor$ rounds x down to the nearest integer. These percentages describe observed reference length, not online checkpoints or known fractions of an uncapped reasoning process.

The main problem set consists of MATH-500 indices 50–149, in dataset order, for both Qwen3-8B modes (Yang et al., 2025a; Hendrycks et al., 2021). The reference cap is 4096 tokens; 38/100 Think and 1/100 NoThink traces are capped. We keep all of them in the main analysis and report the naturally ended and capped groups separately. Selecting only correct or naturally ended traces would change which problems are studied.

At each checkpoint, we test the problem's *own* prefix, the question with *no trace* (question-only, or null in tables), and a *mismatched* prefix from the next problem's reference trace. A shorter mismatched source trace is truncated before its final token, so mismatched and own prefixes need not have identical lengths. This control tests the effect of irrelevant reasoning context. It cannot separate content effects from length effects, because prefix lengths can differ.

3.2 Matched Readouts and Answer Endpoints

We use two continuation conditions, also called arms in tables. The free condition continues the prompt and prefix without an extra request. The forced condition appends the text suffix s_{box},

$$s_{\text{box}} = \text{"\textbackslash nTherefore, the final answer is \textbackslash boxed\{"}. \tag{1}$$

Here `\n` is a newline and `\boxed{` opens a LaTeX answer box. Each condition uses $N = 8$ samples, temperature 1, the same EOS stopping rule, and the same new-token cap B_f. This cap is twice the observed reference length remaining after the prefix, with a minimum of 256 and maximum of 4096 tokens:

$$B_f = \min(4096, \max(256, 2(T_{\text{obs}} - k_f))) . \tag{2}$$

Matching these settings lets us test the added answer request. It does not equalize actual output length, computation, or wall time. Here forced means that an answer is requested, not that generation stops when it appears. The original request in follow-ups is this same suffix.

We score the *first* and *last* complete boxed answers separately. In a forced output, the first answer is the first complete LaTeX box with matching braces whose opening delimiter was supplied by the suffix, not the first closing brace in raw text. The last answer may follow later revisions. For free continuation, first and last refer to complete boxes emitted in the new output; labels such as free-last name the condition and endpoint. We never attach the full prefix during extraction: otherwise an old prefix answer could be mistaken for a new probe answer. If a box begins in the prefix and only closes in a free continuation, it does not count as a newly emitted complete box.

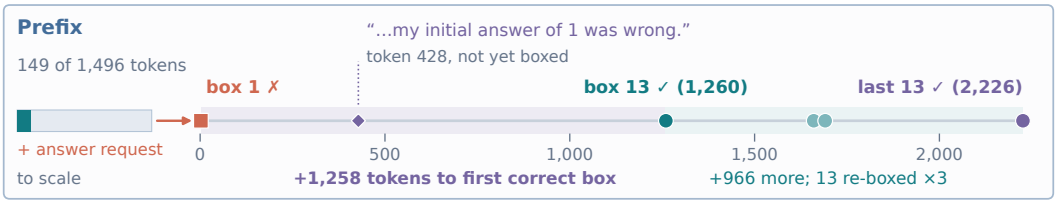

Figure 3: **A saved revision.** For the maximum of $\gcd(n+7, 2n+1)$, with positive integer n, Qwen Think changes its boxed answer from 1 to 13. Token axis to scale; markers show each closed box; the diamond marks the unboxed rejection of 1. Selection and data sources: Appendix K.

For each readout condition, let J count problems and N count samples per problem. At endpoint $e \in \{\text{first}, \text{last}\}$, A^e_{ji} is the answer from sample i for problem j, whose gold answer is a^*_j. The grader $g(A, a^*)$ returns 1 for a correct candidate A against gold answer a^*, and 0 otherwise. Mean per-sample correctness is

$$\widehat{R}_e = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{N}\sum_{i=1}^{N} g(A^e_{ji}, a^*_j). \tag{3}$$

A missing answer counts as unsuccessful recovery, but is reported separately through *answer coverage*: the fraction of samples with an extracted candidate. A complete box containing a placeholder counts toward coverage but not correctness. We score each continuation individually, rather than whether at least one of the eight answers is correct or a majority vote is correct. MATH grading normalizes answer strings and tests mathematical equality for permitted expressions; Appendix D specifies the version and its limits.

What changes after the first answer? We classify each endpoint as correct (C), extracted but incorrect (W), or missing under the extraction rule (M). For $U, V \in \{C, W, M\}$, let p_{UV} be the fraction of all samples whose first state is U and last state is V. For the forced condition, the accuracy change is exactly

$$\widehat{R}_{\text{last}} - \widehat{R}_{\text{first}} = p_{WC} + p_{MC} - p_{CW} - p_{CM}. \tag{4}$$

This accounting separates answer replacement from newly available labels; W includes MATH placeholders, so a changed answer need not mean a corrected solution. Consequently, comparing free-last with forced-first mixes two quantities: the last-answer difference between protocols and the first-to-last change within the forced output. For outputs ending correct after an unsuccessful first answer, we also locate the first later correct box and count generated tokens since the first box closed. These correction times use gold answers and are measured only offline. Nested boxes are not later revisions. The analysis was designed after the endpoint results were known (post-hoc), preserves the original first/last scores, and reports how many outputs lack a measurable delay (Appendix K).

3.3 Controls and Replications

Three follow-ups test whether revision survives changes to the request and reference length. Their designs were fixed after the main analysis but before each follow-up's generation. First, 50 separate MATH problems with new Qwen Think references compare the original request, a shorter natural request, and a request preceded by `</think>` at 10%. This marker closes Qwen's designated reasoning channel (Appendix L).

Second, the *close-reasoning* extension reuses 100 Qwen Think and 50 GPT-OSS MATH references at both checkpoints. It compares newly sampled original requests with requests that close the reasoning channel before asking for an answer. GPT-OSS switches from its `analysis` channel to `final`; Qwen uses `</think>`. Both conditions use eight samples, temperature 1, and cap B_f. Third, 50 reused Qwen questions receive new references with a 16,384-token allowance and 8192-token continuations, checking revision with longer references (Appendix I).

The NoThink budget control combines free/forced continuation, greedy/sampled decoding, and 64-token/B_f caps at 10% on the same 100 prefixes. Greedy selects the highest-probability next token at temperature 0, once per condition; sampled uses eight outputs at temperature 1. Both report per-sample accuracy, not an equal-cost voting decision. At 50%, own-prefix and question-only free continuations receive a common 4096-token cap. An exploratory Think run raises it to 8192 because many question-only outputs still hit 4096. Cap-hit rates show whether the larger allowance is enough for outputs to finish.

The MATH replication uses GPT-OSS-120B with medium reasoning (Agarwal et al., 2025) on indices 200–249, with 8192-token reference and continuation caps. Each problem receives eight free and eight forced outputs per checkpoint; both checkpoints share eight question-only outputs. We score raw text, which decodes all new tokens, in both conditions. Rendered text is the user-visible response and may omit reasoning; Appendix E compares this alternative extraction.

For the science benchmark GPQA-Diamond (Rein et al., 2023), Qwen3-8B NoThink and GPT-OSS-120B share a seeded 100-question subset. Both score raw text and use 8192-token reference and continuation caps, first-sampled references, and eight separately seeded calls per condition, with shared question-only baselines. The MATH replication uses the same sampling procedure; earlier batches with excessive duplication are excluded (Appendix H). Strict extraction requires a boxed option letter. The secondary Label+ rule also accepts explicit labels such as `B) 0.85` inside the selected box; it was defined after inspecting earlier outputs and fixed before resampling (Appendix F).

3.4 Uncertainty and Online Evaluation

We treat each problem, not each continuation, as one independent unit in the statistical analysis. Paired bootstrap resampling draws problems with replacement, keeping their conditions together (Dror et al., 2018). We use 20,000 resamples and use the 2.5th and 97.5th percentiles for 95% confidence intervals (CIs). Differences are in percentage points (pp); n in displays counts problems. Sign-flip tests randomly reverse the sign of each problem-level difference. Their p-value is the fraction of randomized means at least as extreme as the observed mean. Reported intervals are exploratory and do not account for examining many comparisons. The readout study uses one reference per problem; the separate agreement study in Appendix N uses eight. Earlier pilot results are kept separate.

At each checkpoint, the online rule stops only if all eight probes' last boxed answers match in user-visible text after repeated whitespace is collapsed. Missing answers prevent stopping; agreement does not ensure correctness. Checkpoints are fixed main-generation token counts because final length

is unknown online. This tests whether agreement saves work, which offline accuracy alone cannot show.

4 RESULTS

4.1 THE SCORED ANSWER CHANGES THE COMPARISON

Matching answer endpoints reverses the main Think comparison. At 10%, free-last accuracy is 69.3%, while forced accuracy rises from 31.8% at the first answer to 73.8% at the last. The apparent 37.5-point free advantage over forced-first combines a -4.5-point last-answer difference with a 42.0-point gain inside forced outputs. This accounting asks how much of the score arrives after the requested answer; it does not locate a committed answer in the prefix.

The saved revision in Figure 3 also appears across many outputs: 339 of 800 Think outputs change from an unsuccessful first answer to a correct last answer. Of these, 286 begin with a wrong answer other than the explicit placeholders `?`, `...`, or `something`; the remaining 53 begin with one of those placeholders. Median delays are 1,210 tokens for non-placeholder answers and 1,352 for placeholders (Appendix Table 21). Thus placeholder completion alone does not explain the gain.

GPT-OSS-120B also checks the problem and first answer without seeing later answers. Majority judgments classify 89 of 96 distinct problem–answer pairs as concrete attempts, covering 273 of the 286 non-placeholder improvements. This checks form, not correctness, and needs human validation (Appendix K). GPQA gains also include newly recognizable option labels. Thus final-answer accuracy mixes several kinds of progress.

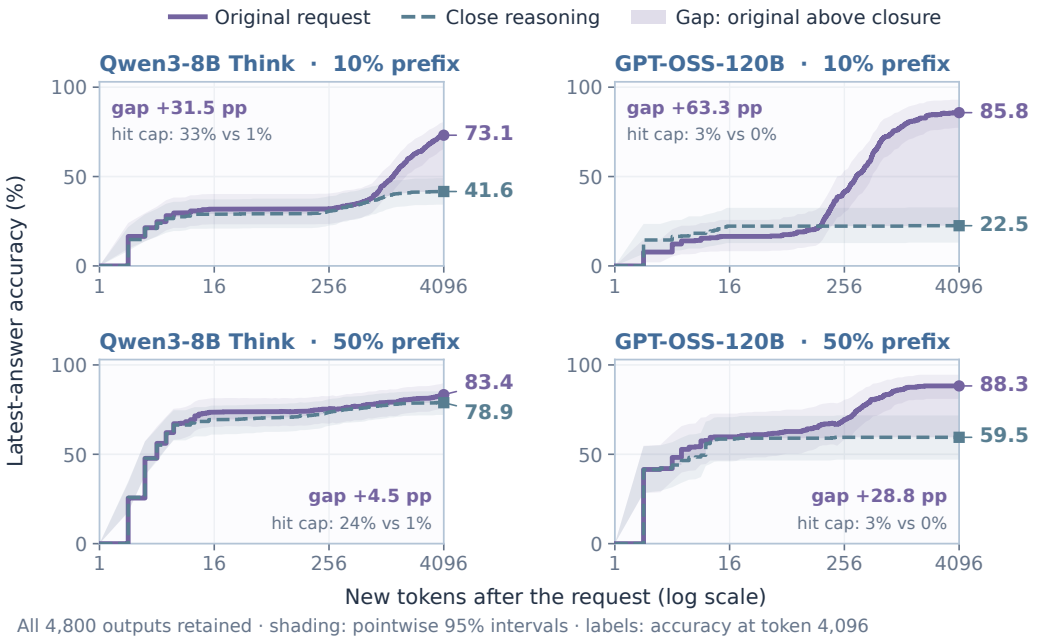

Figure 4: **Later tokens account for much of the recovery.** All 4,800 channel-closure continuations (Table 1), scored by the latest complete box; missing answers count as incorrect, and finished or capped outputs retain their last answer. These are saved trajectories, not fresh runs at shorter budgets. Light bands: pointwise 95% question-bootstrap intervals; darker fill: original above closure. Labels: accuracy at token 4,096. Reconstruction: Appendix L.

Varying the request tests the source of later gain. On 50 separate problems, original and natural requests gain 41.8 and 43.3 points; closing Qwen's reasoning channel reduces the gain to 10.5 (Appendix L). The cross-model extension then tests both checkpoints using new outputs from saved references (Table 1).

The curves show when the later gain arrives (Figure 4). For Qwen at 10%, original-request accuracy is 31.6% after 64 new tokens and 62.8% after 2,048. Closing reasoning limits this later rise; for GPT-OSS, it leaves the curves almost flat after the initial answer. These are changes along the newly sampled follow-up outputs, whose endpoints can differ from the original audit above.

Table 1: **Closing reasoning usually reduces later gain.** Eight new continuations per MATH problem, request, and checkpoint. O/C means Original / Close-reasoning. First is accuracy (%); Gain is last minus first accuracy (pp); Tokens is mean output length. ΔGain subtracts Close gain from Original gain, with a 95% question-bootstrap interval. Positive values mean less gain after closure.

Model	Prefix	First O/C	Gain O/C	ΔGain [95% CI]	Tokens O/C
Qwen3-8B	10%	31.6/29.0	41.5/12.6	28.9 [20.4, 37.3]	2461/255
Qwen3-8B	50%	72.9/69.3	10.5/9.6	0.9 [-2.6, 4.6]	1834/288
GPT-OSS-120B	10%	16.5/22.3	69.3/0.3	69.0 [58.0, 79.5]	685/9
GPT-OSS-120B	50%	61.0/59.0	27.3/0.5	26.8 [17.0, 37.5]	361/7

Channel closure usually reduces post-answer gain, but its effect depends on the model and checkpoint (Table 1). At 10%, it removes 28.9 points of Qwen's gain. For GPT-OSS it removes 69.0 and 26.8 points at 10% and 50%. Qwen at 50% is the exception: both conditions still gain about 10 points, and their 0.9-point difference is uncertain. Closure has smaller effects on first-answer accuracy (Appendix L). It is therefore a useful change to the readout, but does not generally extract an answer without further generation.

4.2 Output Budget Can Reverse the Ordering

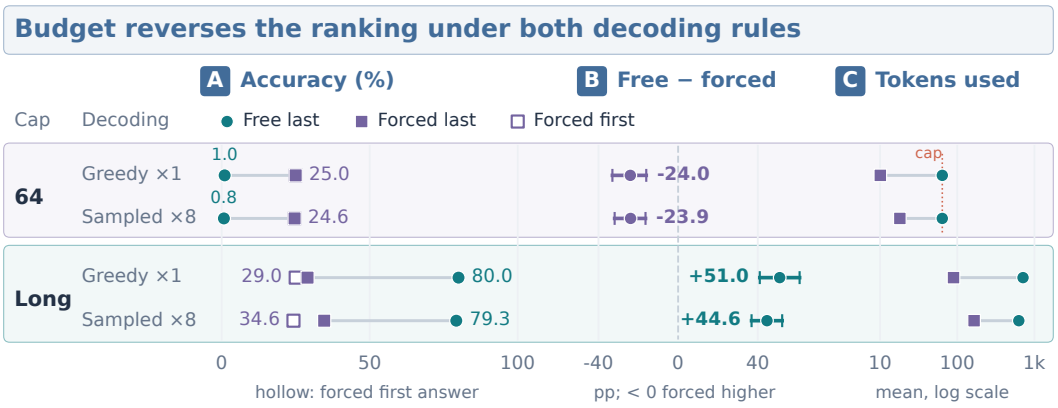

Figure 5: **The larger budget reverses the ordering.** Qwen3-8B NoThink, 100 MATH problems, 10% prefix. (A) Mean last-answer accuracy; connectors pair the protocols. Hollow: forced first answer. (B) Their paired difference with 95% question-bootstrap intervals. (C) Mean generated tokens. Greedy: one output at temperature 0; sampled: eight at temperature 1. Long denotes the reference-dependent cap B_f. Appendix Table 4 adds first-answer accuracy, coverage, and token counts.

The NoThink control shows that changing the output budget can reverse the ordering. Under a 64-token cap, free continuation rarely reaches an answer, whereas the forced request gets one within that limit; under the longer cap, free continuation is more accurate for both decoding rules (Figure 5). Changing the cap shifts the sampled difference between free-last and forced-last by 68.5 points, far more than changing greedy to sampled decoding. The amount of generation available is therefore the main factor in this control.

Equal requested caps do not imply equal generated tokens or time: the suffix changes when answers appear and when generation stops. This control isolates requested budget and decoding settings; efficiency per token or per second needs a separate comparison.

4.3 Missing Answers and Truncated References Matter

Missing answers and reference length affect how we interpret the Think comparison. A failed readout can mean a wrong answer or no extractable answer within the cap. At 10%, only 73.1% of Think free samples produce an extractable answer, versus 100% of forced samples (Appendix Figure 7). Overall accuracy therefore mixes answer quality with answer availability. We report coverage separately and keep unanswered samples in the accuracy denominator, so neither condition is evaluated only on its answered subset.

The last-answer ranking also differs between naturally completed and truncated reference traces. For naturally ended Think references, free-last and forced-last are nearly tied at 10% (+0.8 points); for capped references, free-last is 13.2 points lower. The reversal is therefore concentrated in longer or harder traces, rather than a uniform advantage for forced continuation. The groups also differ in difficulty, so this comparison cannot isolate the effect of truncation.

Longer references preserve revision. All 50 new Qwen Think references end naturally under a 16,384-token limit. This follow-up reuses questions and gives continuations a common 8192-token cap (Appendix I). At 10%, forced first-to-last gain is 52.0 [39.2, 64.2] points, but free-last minus forced-last is only +0.5 [-1.5, 2.5]. Closure reduces gain by 41.2 [28.7, 53.2] points at 10%, versus 2.2 [-4.0, 9.2] at 50%. Thus capped references are not necessary for substantial revision, although this follow-up changes both references and continuation caps. The evidence supports later gain more broadly than the original ranking reversal.

4.4 Does the Prefix Help Beyond the Question?

Table 2: **Apparent prefix benefit depends on the budget.** Free continuations, 50% prefix, 100 MATH problems per mode, eight samples per condition. Own retains the prefix; Question only receives no trace. Gain is Own minus Question only (pp; paired 95% intervals). Capped is the fraction of question-only outputs reaching the limit. B_f varies with reference length; 4096/8192 are common caps. The 8192 run is exploratory.

Mode	Cap	Own	Question only	Gain [95% CI]	Capped
Think	B_f	76.6%	53.5%	23.1 [18.1, 28.3]	70.3%
Think	4096	75.4%	66.4%	9.0 [5.6, 12.8]	37.1%
Think	8192	88.9%	83.9%	5.0 [2.1, 8.5]	15.8%
NoThink	B_f	79.5%	56.8%	22.8 [16.6, 29.1]	38.5%
NoThink	4096	80.4%	81.9%	-1.5 [-4.9, 1.4]	0.8%

After accounting for later generation, we ask whether the prefix helps beyond the question. At 10%, own-prefix minus question-only accuracy is small and inconsistent across modes; Think's bootstrap interval and sign-flip test disagree in this borderline case. Mismatched prefixes reduce free-continuation accuracy (Appendix Table 6), but interference from an irrelevant trace does not show that the correct trace is needed. Early recovery alone therefore does not show that the answer was already available in the prefix.

At 50%, the default cap makes the prefix appear more helpful because question-only continuations often run out of tokens (Table 2). In NoThink, the apparent gain is 22.8 points when 38.5% of question-only outputs hit their cap. With a common 4096-token allowance, that cap-hit rate falls to 0.8% and the difference becomes -1.5 [-4.9, 1.4] points. The readout budget therefore changes whether this experiment detects a prefix advantage.

Think retains a positive benefit as the allowance increases: +9.0 [5.6, 12.8] points at 4096 and +5.0 [2.1, 8.5] at 8192. Even at 8192, 15.8% of question-only outputs reach the cap, so this is not a comparison in which every continuation has finished. The reduced gap is consistent with the prefix leaving less work to do; the remaining positive gap is evidence of benefit under these budgets, without identifying its mechanism.

A trace can carry answer-specific information, reduce remaining computation, or do both. Equal *additional* budgets cannot separate these roles: only the own-prefix arm has received a partial solution. Think's positive gains are compatible with either role; NoThink's interval spanning zero establishes neither equivalence nor the absence of useful answer-specific information.

4.5 Replication Across Models and Tasks

GPT-OSS also shows that the scored endpoint changes the comparison, while its prefix benefit remains uncertain. At 10% on MATH, GPT-OSS's free-last advantage falls from 76.8 points against forced-first to 7.5 against forced-last, while own-prefix and question-only continuations show no clear accuracy difference (Appendix E).

On four-option GPQA, random choice has 25% accuracy. Under the strict extraction rule, Qwen reaches 26.1–36.9% accuracy, with 28–42% of labels missing. These results mainly show the effect of missing or differently formatted answers and provide limited evidence about prefix value. Its 10% free-last versus forced-first contrast changes from -2.6 to $+11.5$ points when extraction

accepts explicit option labels, because recognized free-answer coverage rises from 71.8% to 97.1% (Appendix F). The generations are unchanged; extraction alone changes the measured comparison. At 10% on GPQA, GPT-OSS again shows a large free-last advantage that shrinks after forced revision (+46.9 to +7.3 points). Fixed-cap own-prefix comparisons do not show a clear advantage over the question-only baseline for either tested model, although Qwen's near-chance accuracy under the strict rule limits that inference. The GPQA results extend the endpoint and formatting sensitivity to science questions. They provide limited evidence about what answer information the prefix contains.

4.6 Does Online Probing Save Time?

The final test concerns deployment: even useful readouts must save more work than probing adds. On 100 additional NoThink MATH problems, the main generation pauses after 256, 512, and 1024 tokens to launch eight free probes. It stops early if all eight answers agree. It uses 31% fewer main tokens but 2.59× the wall time and 8.92× total generation (Figure 1C). One-shot, checkpoint-only, and online means are 5.3, 7.3, and 13.8 seconds, respectively. Here checkpoint-only pauses and resumes without probing. Four of 58 unanimous stops give wrong answers. After the answer-format audit, accuracy is 83% online versus 88% one-shot. The probe cost outweighs the saved main-generation work, so this rule provides no wall-clock speedup. Appendix G reports the full accounting. This result applies to the tested synchronous rule; cheaper probes and other parallel implementations require separate evaluation.

5 Discussion and Limitations

What to report and compare. An answer-recovery result should specify the request and channel transition, reference termination, new-token budget, answer endpoint, and extraction rule. Report first and last accuracy together with coverage and post-answer tokens; use a question-only control before attributing gains to the prefix. These measurements distinguish immediate answering from later generation, without identifying an internal decision or mechanism. They do not invalidate published short-probe results whose protocol differs from ours.

A useful stopping policy needs its own validation. Our untuned rule can stop on wrong unanimous answers and did not improve latency in our audit. A competitive policy needs error measurement on separate data and equal-resource latency comparisons, including early-exit baselines.

Scope. Evidence covers two model families and two benchmarks. Separate MATH problem sets support the main study, replication, and online audit; some controls reuse them. Matching text does not guarantee matching meaning, and GPQA parsing can miss answers that do not state an option label. The three-request follow-up covers one model and checkpoint; the closure extension covers two models and checkpoints but only MATH and reuses saved references. The main Qwen reference cap is 4096 tokens, below the 32,768-token output allowance recommended by the model authors (Qwen Team, 2025); 38% of Think references are capped. The 16k follow-up has no capped references, but reuses questions and changes continuation caps; it does not validate the 32k regime. Fractional checkpoint positions also move. Broader models and independent human grading review are needed.

6 Conclusion

A correct continuation alone does not show readiness to stop. Endpoint audits reveal substantial later gains even with naturally ended, longer references. Channel closure reduces these gains in several settings, with model and checkpoint exceptions. Readout accuracy must therefore be interpreted alongside the generation that produces it.

Prefix benefit and stopping efficiency require separate evidence. Question-only controls retain positive Think gains under common caps, while our online rule saves main tokens but increases time and total generation. Together, the audits separate answer recovery, prefix benefit, and useful stopping. Report evidence for each claim; an effective exit policy still requires error validation and equal-resource latency comparisons.

Reproducibility Statement

Artifacts include raw tokens, prompts, configurations, seeds, versioned graders, paired analyses, and figure/table builders. File fingerprints preserve the record of changes; Appendix H documents sampling repair and which saved files feed the current results. Every controlled table is generated from audited outputs. Recorded API and package versions aid reproduction but cannot guarantee exact replay of outputs from a hosted service.

Ethics Statement

We evaluate public benchmarks without human subjects. The main practical risk is premature deployment of unreliable stopping rules.

Artificial Intelligence (AI) Use Statement

We used Codex to help draft, translate, and polish the manuscript, and for literature retrieval and discovery, including finding and comparing related work. Additional assistance covered refining hypotheses and arguments, experimental design, writing and testing code, executing API experiments, analyzing results, and preparing figures and tables. AI-assisted correctness review and the blinded cross-model content audit are described in the relevant protocols. Benchmark questions were not AI-generated. Automated checks do not replace author review; the authors remain responsible for all claims, citations, code, and experimental results.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

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

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv preprint arXiv:2603.05488*, 2026.

Yigit Utku Bulut. It's the problem, not the path: Budget and difficulty confounds in llm reasoning trajectories. *arXiv preprint arXiv:2609.03436*, 2026.

K Cox, D Kianersi, and A Garriga-Alonso. Post-hoc reasoning in chain of thought: Decoding and steering pre-committed answers. *arXiv preprint arXiv:2603.01437*, 2603, 2026.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, et al. Efficiently scaling llm reasoning with certaindex. *arXiv preprint arXiv:2412.20993*, 2024.

Yichao Fu, Junda Chen, Yonghao Zhuang, Zheyu Fu, Ion Stoica, and Hao Zhang. Reasoning without self-doubt: More efficient chain-of-thought through certainty probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Naoto Iwase, Yuki Ichihara, Mohammad Atif Quamar, and Junpei Komiyama. Reliable chain-of-thought via prefix consistency. *arXiv preprint arXiv:2605.07654*, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Yunxiang Mo, Donghao Zhao, and Hejia Geng. Stable answers, unfinished reasoning: Why self-consensus is not a safe early-exit signal. *arXiv preprint arXiv:2609.09989*, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Qwen Team. Qwen3-8b model card, 2025. URL https://huggingface.co/Qwen/Qwen3-8B. Best Practices, accessed September 21, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 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. *arXiv preprint arXiv:2203.11171*, 2022.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examining the necessity and accuracy of cot step-by-step. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2586–2606, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. *arXiv preprint arXiv:2504.15895*, 2025b.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

A Protocol and Statistical Details

Uncertainty. For each contrast, we first compute the difference between the relevant sample averages for each problem. We resample these differences 20,000 times and report percentile intervals, keeping each problem's outputs together. Sign-flip tests use 100,000 draws. The two procedures need not agree on borderline comparisons. Intervals are not adjusted for multiple comparisons and exclude variation from new reference draws or API versions. The single-problem censored group receives no confidence interval.

Table 3: Controlled experiment inventory. All reference traces are first-sampled without label selection. MATH problem sets use original dataset indices; GPQA uses a seeded subset. Own/null means free continuation unless stated otherwise. N counts samples per problem; B_f is the fraction-dependent probe cap from Section 3.2.

Experiment	Settings
Qwen matched	100 problems per mode (50–149); reference cap 4096; own/null/mismatched at 10% and 50%; free and forced $N = 8$, temperature 1, cap B_f, EOS stops.
NoThink factorial	Same 100 prefixes at 10%; free/forced $\times$ greedy/sampled $\times$ 64/B_f caps. Greedy: $N = 1$, temperature 0; sampled: $N = 8$, temperature 1.
Fixed-cap prefix	Same 100 problems per mode at 50%; own/null, $N = 8$, temperature 1, common cap 4096; exploratory Think cap 8192.
GPT-OSS replication	50 problems (200–249); medium reasoning; reference and probe cap 8192; own/free and own/forced at 10% and 50%; shared null/free baseline; $N = 8$, temperature 1, EOS stops.
GPQA replication	Same seeded 100 of 198 questions in Qwen3-8B NoThink and GPT-OSS-120B; first reference, common 8192-token caps; own/free and own/forced at 10% and 50%, shared null/free; $N = 8$, temperature 1.
Suffix follow-up	50 separate MATH problems (0–49); Qwen3-8B Think; new references capped at 4096; 10% checkpoint; original, natural, and end-think requests; eight distinct single-sample calls per arm.
Online audit	100 problems (300–399); NoThink; main cap 4096; checkpoints 256/512/1024; eight concurrent free probes capped at 2048; exact normalized 8/8 vote; one problem evaluated at a time.

Prompt and stop rules. For MATH, the benchmark software supplies worked examples (a few-shot prompt) and the question. GPQA uses no demonstrations (Appendix F). A model-specific renderer converts these messages into model input. The forced condition appends its request inside the ongoing completion, not as a new user turn. Matched and factorial samples stop at EOS or the cap; forced-first is read afterward. Qwen free readout uses rendered text; GPT-OSS retains both raw text and text from only the rendered final-response section because channel handling can change extraction.

Problem sets and reuse. Factorial and fixed-cap experiments reuse reference prefixes but draw new continuations; their scores need not equal similarly named conditions in the main matched-readout grid. The earlier 50-problem design pilot is excluded from the main problem set. The historical online pilot is exploratory, and index 150 had an engineering check. The new 100-problem online set shares no problems with that pilot or the controlled MATH studies. The second-family sample size, caps, checkpoints, and contrasts were fixed before querying that model; no calibration on separate data is claimed.

B Complete Controlled Readouts

Table 4: **Budget, suffix, and decoding controls.** NoThink, 100 MATH problems, 10% checkpoint. Long means B_f; sampled means eight outputs at temperature 1 and greedy means one at temperature 0. Correctness is per-sample. Free coverage records whether an answer was extracted; all forced cells have 100% coverage. Tokens are means, free / forced.

Decoding rule	Cap	Free last	Forced first	Forced last	Free coverage	Tokens
Greedy	64	1.0%	25.0%	25.0%	1.0%	64 / 10
Greedy	Long	80.0%	25.0%	29.0%	90.0%	711 / 89
Sampled	64	0.8%	24.6%	24.6%	0.8%	64 / 18
Sampled	Long	79.3%	24.3%	34.6%	96.5%	627 / 165

C Reference-Length Censoring

The Qwen reference cap truncates 38 Think and one NoThink reference trace. All subsequent fractions and dynamic probe caps use the observed trace length. Table 7 preserves these problems in the main estimate and shows the separate groups. A capped reference is likely to be a harder or longer

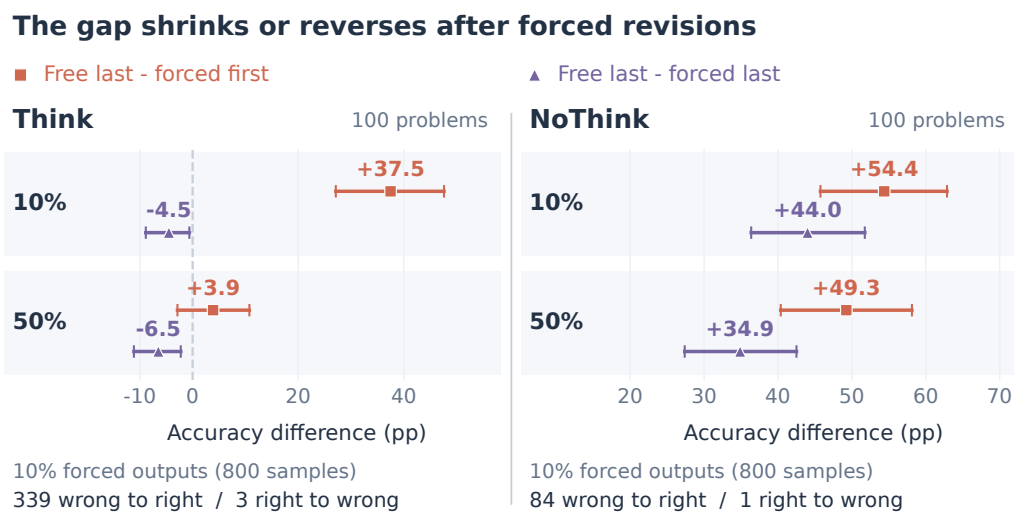

Figure 6: **The endpoint changes the comparison of free and forced continuation.** Paired accuracy differences on 100 Qwen3-8B problems per mode, with 95% problem-bootstrap intervals. Positive values favor free-last. Fractions use observed reference lengths, including capped traces. Free-last scores the last new box without a request; forced-first/last score the same requested output's first/last box. Axes differ to show within-mode contrasts.

Table 5: First/last endpoint audit on the matched Qwen problem set. All entries are per-sample accuracy; missing answers are unsuccessful recovery. Free first and last use newly emitted boxes only. Forced first is specifically the box opened by the request.

Mode	Prefix	Free first	Free last	Forced first	Forced last
Think	10%	67.8%	69.3%	31.8%	73.8%
Think	50%	75.5%	76.6%	72.8%	83.1%
NoThink	10%	77.9%	78.8%	24.4%	34.8%
NoThink	50%	79.1%	79.5%	30.3%	44.6%

Table 6: Full own-prefix / question-only (null) / mismatched-prefix grid, $n = 100$ per mode. Free and forced columns score the last answer; the gap is free minus forced, in percentage points, with paired 95% intervals. "Prefix" is the reference fraction setting the cap; null receives no trace. Requested caps and sample counts match within every pair. Mismatched prefixes are not always length-matched.

Mode	Prefix	Context	Free last	Forced last	Gap [95% CI]
Think	10%	own	69.3%	73.8%	-4.5 [-8.9, -0.6]
Think	10%	null	66.5%	69.6%	-3.1 [-10.4, 4.1]
Think	10%	mismatched	7.6%	38.1%	-30.5 [-37.3, -24.0]
Think	50%	own	76.6%	83.1%	-6.5 [-11.1, -2.3]
Think	50%	null	53.5%	67.6%	-14.1 [-21.4, -7.0]
Think	50%	mismatched	2.3%	14.1%	-11.9 [-16.6, -7.6]
NoThink	10%	own	78.8%	34.8%	44.0 [36.4, 51.5]
NoThink	10%	null	80.0%	38.1%	41.9 [35.1, 48.5]
NoThink	10%	mismatched	4.6%	12.1%	-7.5 [-11.3, -4.3]
NoThink	50%	own	79.5%	44.6%	34.9 [27.4, 42.6]
NoThink	50%	null	56.8%	32.4%	24.4 [17.1, 31.8]
NoThink	50%	mismatched	0.0%	3.4%	-3.4 [-6.6, -1.0]

problem, so the groups cannot isolate what cutting off the reference alone would change. Increasing a reference cap would also change prefix positions and is a different experiment.

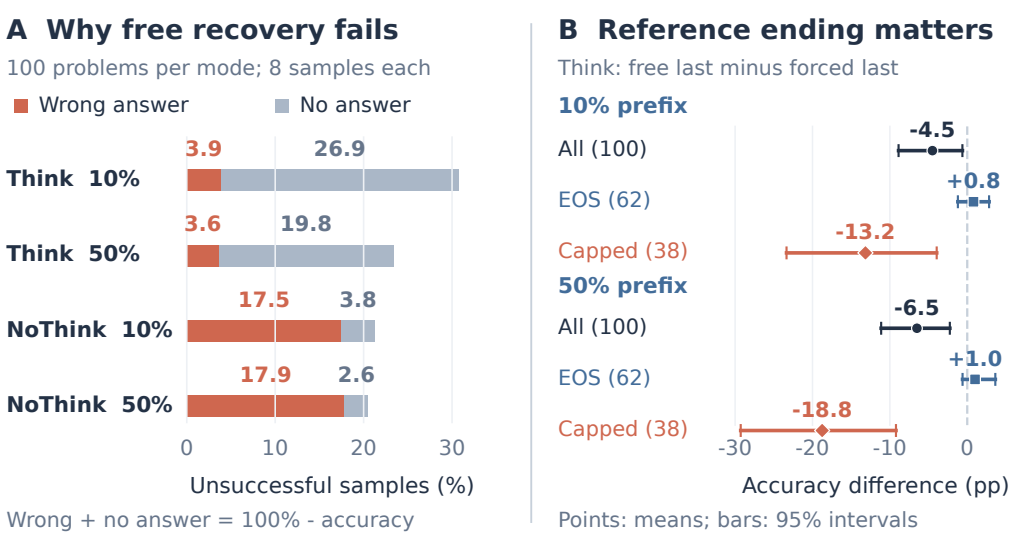

Figure 7: **Two checks on the overall result.** (A) Failed free continuations split into extracted wrong answers and missing answers, as percentages of all samples. Coverage equals 100% minus the missing share. (B) Think's free-last minus forced-last difference at both checkpoints. Points/bars show paired problem means/95% bootstrap intervals; pp means percentage points. EOS is natural end-of-sequence; capped means the reference hit its token limit, and parentheses count problems. The groups differ in difficulty: this does not estimate what truncation alone would change. Full groups appear in Table 7.

Table 7: Last-answer accuracy by reference termination. "Censored" means the reference ended at its cap without EOS; n counts problems in each row. Differences are free minus forced in percentage points. The single censored NoThink case is descriptive only; its interval is omitted.

Mode	Prefix	Group	n	Free	Forced	Gap [95% CI]
Think	10%	all	100	69.3%	73.8%	-4.5 [-8.9, -0.6]
Think	10%	eos	62	94.6%	93.8%	0.8 [-1.2, 2.8]
Think	10%	censored	38	28.0%	41.1%	-13.2 [-23.4, -3.9]
Think	50%	all	100	76.6%	83.1%	-6.5 [-11.1, -2.3]
Think	50%	eos	62	96.2%	95.2%	1.0 [-0.6, 3.6]
Think	50%	censored	38	44.7%	63.5%	-18.8 [-29.3, -9.2]
NoThink	10%	all	100	78.8%	34.8%	44.0 [36.3, 51.8]
NoThink	10%	eos	99	78.7%	34.7%	43.9 [36.0, 51.8]
NoThink	10%	censored	1	87.5%	37.5%	–
NoThink	50%	all	100	79.5%	44.6%	34.9 [27.3, 42.5]
NoThink	50%	eos	99	79.4%	44.2%	35.2 [27.7, 42.8]
NoThink	50%	censored	1	87.5%	87.5%	–

D Extraction and Grading Audit

Local box parsing. A prefix can end inside a LaTeX expression. The newly generated suffix may then begin with unmatched closing braces even though a valid final `\boxed{...}` appears later. The original parser raised an error on an unrelated unmatched brace and lost that final answer. The corrected parser searches each box locally, balances nested braces, treats escaped braces as literals, and chooses the last complete box. An incomplete later box does not invalidate an earlier complete one. The legacy unbraced `\boxed 12` fallback is retained. No gold answer is used to select a candidate.

Symbolic guard and conservative matching. The original input check reversed its intended rule: it required patterns that should have been rejected. This prevented ordinary expressions from reaching symbolic comparison. The repair rejects those patterns, restricts inputs to simple arithmetic syntax, and bounds grading time. Tests cover equivalent fractions and decimals, radicals,

Table 8: Effect of the grading repair on own/free outputs. "Original" reproduces cached scores; "Parser" changes extraction only; "v2" is the repaired grader version, also fixing the input-safety check for symbolic comparison (symbolic guard). Coverage is measured after the repair. The paired changes are not new model samples.

Mode	Prefix	Original	Parser	v2	Coverage
Think	10%	69.1%	69.3%	69.3%	73.1%
Think	50%	76.6%	76.6%	76.6%	80.3%
NoThink	10%	71.5%	78.8%	78.8%	96.3%
NoThink	50%	73.5%	79.5%	79.5%	97.4%

algebraic expressions, incorrect answers, malformed boxes, escaped braces, and unsafe or expensive expressions. Some older conservative policies remain: ordered tuples require matching boundaries, and certain integer-versus-expression and unreduced-fraction pairs are not accepted. This does not recognize all mathematically equivalent answers.

Reproducible correction. The v2 auditor first reproduces every saved answer and correctness count with a saved copy of the original grader. It then re-extracts from raw tokens, separately records parser-only and symbolic changes, and writes new result files with one record per line, plus input hashes, without modifying the originals. All active Qwen controlled tables and plots use the repaired outputs. The historical 20-problem online pilot ran with the old parser: only its returned answers were regraded, preserving its votes and timings. The current 100-problem online audit instead executes the corrected parser live.

Raw-versus-rendered sensitivity. We also score both endpoints from raw newly generated text across all 27,600 Qwen samples. Relative to the original readout of rendered free continuations, 43 of 55,200 correctness decisions change, all at null/free *first* endpoints. Every own-prefix score and every last-answer score used in the main comparisons is unchanged. This check does not resample the models.

E Second-Family Replication Details

The GPT-OSS-120B run saves 50 first reference samples and 2000 continuation samples: four own-prefix arms and a shared null arm, each with eight samples per problem. The references are unchanged from the initial run; all reported continuations were resampled in separate calls with distinct seeds. It retains token sequences, seeds, stop reasons, reference prompts, and both raw and rendered readouts. Reference and probe caps are 8192. The shared null baseline avoids independently resampling an identical condition for each fraction; paired comparisons retain the resulting dependence at the problem level. The protocol and exact software versions are stored alongside the data. Request times are logged for record keeping, not used as controlled latency measurements.

Table 9: **MATH second-family replication.** GPT-OSS-120B, 50 problems, medium reasoning, eight samples per problem and condition, common 8192-token reference and continuation caps. All references end naturally; eight of 2000 continuations hit the cap. The same null/free samples serve both checkpoints. Gaps and 95% intervals are in percentage points.

Prefix	Free last	Forced first	Forced last	Null free	Own − null [95% CI]
10%	95.8%	19.0%	88.3%	95.8%	0.0 [-1.3, 1.3]
50%	96.3%	58.5%	88.3%	95.8%	0.5 [0.0, 1.3]

Primary results. All 50 references terminate at a model stop marker; the longest has 5653 tokens. Eight continuations reach the 8192-token cap: two null/free, four 10% free, one 10% forced, and one 50% free. New continuations generate 1,460,953 tokens; the reused references add 47,894 previously generated tokens. At 10%, free-last minus forced-first is 76.8 points (95% CI [66.8, 86.0]), versus 7.5 ([2.8, 13.3]) against forced-last. Free-first accuracy is 93.0% at 10% and 95.0% at 50%. Free-first minus forced-first is 74.0 points at 10% (95% CI [63.5, 83.8]) and 36.5 at 50% ([24.3, 48.8]). Thus a large free-versus-forced difference remains even when both conditions are scored at their first complete box.

Grading sensitivity. The fixed primary grader retains conservative integer/fraction rules. We also apply a more permissive scalar-arithmetic comparison to every arm and endpoint. The rule was developed post-hoc on the earlier data and fixed before this repair; the repaired audit covers all 8000 raw/rendered, first/last endpoint decisions on the 2000 new continuations. It accepts unsimplified equivalent expressions and bounded whole-answer binomial notation such as `C(8,2)` (8 choose 2, equal to 28), without changing extraction. The four changed decisions were individually checked: two 10% forced-first answers give `C(8,2)` for 28, and two 50% forced-first answers give $5 - 125 + 5^2 \cdot 5$ for 5. Their raw accuracy becomes 19.5% and 59.0%; free and last-forced scores do not change. In the separately audited, unchanged Qwen own-prefix cells, any endpoint's accuracy changes by at most 1.0 point. The fixed-cap NoThink own/null gap becomes -2.0 points instead of -1.5. These checks retain the main endpoint and budget conclusions; they do not claim complete equivalence checking.

Channel sensitivity. The primary readout scans new raw output symmetrically in both arms, reconstructing only the box explicitly opened by the forced suffix. Parsing only the rendered final-response section can lose that prompt-opened box or omit a box in a reasoning channel. Raw and rendered last-answer strings agree for all free samples; they differ in one forced sample at 10% and none at 50%. These are different extraction procedures applied to the same responses, not additional sampled conditions. The saved records retain both so that channel handling cannot silently change the interpretation.

F Cross-Domain Replication Protocol

Data and prompts. We use a fixed GPQA-Diamond dataset version, revision `633f5ee89ab8a` (full revision in the protocol). A seeded draw selects 100 of its 198 questions without replacement: 44 physics, 49 chemistry, and seven biology questions. The same question indices and independently shuffled option orders are used for Qwen3-8B NoThink and GPT-OSS-120B medium reasoning. The prompt requests reasoning and a final boxed option letter (A, B, C, or D). It contains the question and four options, but no explanation, validator information, gold label, demonstrations, or retrieval. The selected indices, option mappings, tokenized prompts, input hash and software versions accompany the results.

Sampling and endpoints. Each model uses its first sampled reference, an 8192-token cap for both references and every probe, temperature 1, and the model-specific end-of-sequence markers as stopping rules. The references are retained from the original run; each reported probe is a new single-sample call with a distinct seed (Appendix H). At 10% and 50% of observed reference length, it draws eight free and eight forced samples; one eight-sample null/free baseline is shared across checkpoints. Raw newly generated text is scored symmetrically in both arms, reconstructing only the box opened by the forcing request. The first or last box must contain exactly one uppercase option letter after removing whitespace, optional parentheses, or a single `\text` / `\mathrm` wrapper. An invalid box is not skipped in search of a more favorable endpoint. We never search prose for a standalone letter. Missing and invalid answers count as incorrect; valid-letter coverage is reported separately.

Option-label sensitivity. The strict primary readout rejects boxes such as `B) 0.85`, even though they name an option. We developed a secondary rule after examining the original Qwen batch and fixed it before this repair. We apply it unchanged to both models, all arms, and both endpoints. The secondary rule also accepts a leading uppercase option label followed by `)`, `.`, or `:`, or enclosed in parentheses, optionally inside a text-formatting wrapper. A scan of all accepted endpoint boxes finds no multiple explicit option labels. The rule neither searches prose nor matches arbitrary answer content against the gold option; it does not skip an invalid endpoint. This is a post-hoc sensitivity analysis, not a change to the fixed primary measure or new generation.

Coverage, revisions, and channels. All Qwen references end naturally; one GPT-OSS reference reaches the cap. Probe cap hits are rare: 0/4000 for Qwen and 34/4000 for GPT-OSS. Among the latter's 99 EOS references, the early free-last minus forced-first difference remains 47.2 points (CI [38.1, 55.9]); the single censored reference receives no confidence interval. At 10%, 335 GPT-OSS forced samples change from an unsuccessful first readout to a correct last answer: 151 start with a valid wrong letter and 184 with an invalid first endpoint. Eighteen change in reverse. Thus later success includes both corrected choices and newly available option answers. Raw and rendered

Table 10: **Complete GPQA readout audit.** First and last columns are per-sample correctness; coverage requires a valid boxed A/B/C/D at that endpoint. Capped is the fraction ending at the 8192-token limit. Each row contains 800 newly sampled probes. Tokens are mean generated output.

Arm	First	Last	First coverage	Last coverage	Capped	Tokens
Qwen3-8B NoThink ($n = 100$)						
Null free	26.1%	31.3%	57.8%	68.0%	0.0%	1342
10% free	27.8%	32.4%	60.5%	71.8%	0.0%	1242
10% forced	35.0%	36.9%	96.8%	94.5%	0.0%	157
50% free	28.6%	33.4%	59.5%	70.8%	0.0%	771
50% forced	35.0%	36.9%	87.8%	88.4%	0.0%	116
GPT-OSS-120B ($n = 100$)						
Null free	65.6%	67.8%	92.5%	94.5%	1.9%	2322
10% free	66.4%	67.8%	91.5%	93.0%	1.4%	2075
10% forced	20.9%	60.5%	62.5%	94.1%	1.0%	1382
50% free	65.6%	69.3%	90.5%	94.1%	0.0%	1154
50% forced	55.0%	64.5%	90.1%	96.3%	0.0%	648

Table 11: **Paired GPQA endpoint contrasts.** Mean differences and 95% problem-bootstrap intervals, in percentage points. All contrasts retain the same 100 problems, including censored references. These descriptive intervals are not adjusted for multiple comparisons.

Prefix	Free first − forced first	Free last − forced first	Free last − forced last
Qwen3-8B NoThink ($n = 100$)			
10%	-7.3 [-14.8, 0.1]	-2.6 [-10.1, 4.8]	-4.5 [-11.6, 2.6]
50%	-6.4 [-12.9, 0.0]	-1.6 [-8.1, 4.8]	-3.5 [-9.8, 2.8]
GPT-OSS-120B ($n = 100$)			
10%	45.5 [36.6, 54.0]	46.9 [37.8, 55.6]	7.3 [1.2, 13.3]
50%	10.6 [2.9, 18.5]	14.3 [6.7, 22.0]	4.8 [-0.5, 10.3]

Table 12: **Uniform explicit-label sensitivity (Label+).** The same outputs and endpoints as Table 10, now accepting explicit leading option labels. Coverage requires a label recognized by this rule.

Arm	First	Last	First coverage	Last coverage
Qwen3-8B NoThink ($n = 100$)				
Null free	41.4%	47.5%	85.1%	97.4%
10% free	42.1%	47.3%	86.8%	97.1%
10% forced	35.8%	38.6%	99.8%	99.5%
50% free	41.4%	47.5%	85.6%	97.6%
50% forced	38.3%	41.9%	95.9%	99.8%
GPT-OSS-120B ($n = 100$)				
Null free	66.0%	68.1%	93.4%	95.5%
10% free	66.8%	68.3%	92.1%	93.9%
10% forced	20.9%	61.6%	62.5%	95.4%
50% free	66.1%	69.9%	91.0%	94.9%
50% forced	55.0%	64.6%	90.1%	96.5%

last-letter scores agree for all GPQA outputs. Rendered-only parsing loses many prompt-opened first boxes: GPT-OSS forced-first accuracy would be 0.0% / 13.9% at 10% / 50%, instead of the raw-text 20.9% / 55.0%. The raw readout is used in both arms.

Pre-specified design and scope. The question draw, models, option shuffles, sample counts, caps, endpoints and contrasts were fixed before the original calls and retained in the repair. Each model contributes 100 reused references and 4000 new probes. The repair's protocol and runner hashes were saved before generation. GPQA had appeared in the historical survey and the earlier controlled batches; these repaired samples are not a completely unseen benchmark. The extension adds scientific multiple-choice reasoning, not an unrestricted claim across tasks or a controlled model-scaling comparison. Raw/rendered endpoint disagreements, box coverage, and reference EOS/censored groups are retained in the accompanying analysis.

G Online Cost Accounting

Table 13: **Fewer main tokens, greater total cost.** Online NoThink audit ($n = 100$). Accuracy reports primary scoring / audited rescoring of returned answers; the audit corrects seven equivalent-form failures across all 300 answers. Wall time includes API calls and all probe waits. Total tokens sum main and probe output, including failed-to-trigger rounds. Time and token columns are means.

Policy	Accuracy	Wall time	Main tokens	Total tokens
One-shot	85.0% / 88.0%	5.3 s	686	686
Checkpoint-only	84.0% / 86.0%	7.3 s	621	621
Online agreement	81.0% / 83.0%	13.8 s	476	6127

Each reported policy runs on all 100 problems with separate temperature-1 samples. Only probe requests within a checkpoint run in parallel; the main trace waits for the whole probe batch. Total generated tokens equal main tokens plus tokens from every completed probe, including rounds that do not trigger. Auditing all 1462 requests verifies at most eight concurrent requests and no overlap between the main trace and its probes or between policies. All 300 policy records complete without retries, generating 743,461 tokens. All returned answers are extractable and no main trace hits its cap. Fifty-two probes lack an answer; these prevent a trigger. Model-client setup precedes timing and grading follows it; API waiting, parsing, and request logging remain timed. Server load, batching, and hardware are not fixed by this client-side concurrency limit. The 42 cases without a trigger average 19.2 seconds online versus 5.7 seconds one-shot. Because these groups are selected using observed outcomes, their difference does not identify what caused the timing difference.

One-shot and checkpoint-only baselines separate the rough effect of chunking from the added probe work, but their different random traces mean the latency difference is not a deterministic replay of one shared generation. Exact answer-string agreement also misses mathematically equivalent strings. Changing vote normalization would define a new policy requiring new execution.

Table 14: Online policy minus each baseline, $n = 100$. Entries are paired means and 95% problem-bootstrap intervals. Accuracy is evaluated both with the unchanged primary grader and the post-hoc answer-format audit.

Metric	Online − one-shot	Online − checkpoint-only
Accuracy, strict (pp)	-4.0 [-10.0, 2.0]	-3.0 [-9.0, 3.0]
Accuracy, audited (pp)	-5.0 [-11.0, 0.0]	-3.0 [-9.0, 3.0]
Wall time (s)	8.5 [5.6, 12.0]	6.4 [3.5, 10.1]
Main tokens	-210 [-292, -133]	-145 [-231, -63]
Total tokens	5440 [3728, 7351]	5505 [3770, 7494]

Returned-answer format audit. We reviewed all 50 primary grading failures and applied the same secondary rules to all 300 returned answers. Seven failures are equivalent forms: `4343` versus `4343_6` when the question explicitly requests base six (one-shot, index 338); 12th versus 12 when asking for a school grade (all policies, index 379); and $[-2, 7]$ versus $x \in [-2, 7]$ (all policies, index 383). Approximate 37.7 for 12π remains incorrect because rounding was not requested. The primary grader, vote decisions, and timings are unchanged. Five triggers fail the primary grader; the interval-notation case is equivalent, leaving four genuinely wrong unanimous answers. At index 394,

for example, all eight probes vote for $x = -3/4$ in a sequential game: one player chooses real x to minimize the payoff z, then the other chooses real y to maximize it, where $z = (2x - y)^2 - 2y^2 - 3y$. Maximizing over y gives $8(x + 3/8)^2 + 9/8$, so the minimizing player's correct choice is $x = -3/8$. The unanimous value is not an alternative representation of that answer.

Historical pilot. The earlier 20-problem run (indices 150–169) used the legacy parser and had two incomplete attempts restarted on one problem. Its retained one-shot / online means were 18.9 / 45.9 seconds and 947 / 9487 generated tokens, with original-grader accuracy 75% / 80%. Those actual decisions and timings remain archived, not pooled with the current run or relabeled as execution of the corrected policy. Differences in API conditions prevent comparing latency before and after across the two problem sets.

H Sampling Repair and Data Sources

Sampling repair for the reported replications. Initial GPQA and GPT-OSS/MATH batches used one explicit seed with eight requested samples per call and produced extensive exact token-sequence duplication. A controlled interface diagnostic found this pattern with seeded eight-sample calls, but not with separate, differently seeded calls or unseeded batches on the tested prompts. This does not identify the provider's internal random-number implementation. We therefore retained every original question, prompt, and first reference, and regenerated all 10,000 continuation samples in the three affected studies using separate, distinct-seed calls. These replacement samples supply all replication results in this paper. Table 15 summarizes the change; repeated short answers remain valid samples and are not deleted. Distinct token sequences alone do not prove independence. Original batches and before/after analyses remain archived. Since provider conditions may change over time, before/after score changes are not a pure estimate that isolates seeding as the cause. These are repairs, not new confirmations on previously unused data.

Table 15: **Sampling records for the repaired replications.** Groups contain eight responses. Old/new are the mean numbers of distinct token sequences per group, not distinct answer labels. The first two rows use GPQA. Output tokens count all new continuations; original references are retained and charged separately in the saved cost records.

Study	Groups	Old	New	Output tokens
Qwen3-8B NoThink	500	1.63	6.73	2903067
GPT-OSS-120B	500	2.21	7.81	6063999
GPT-OSS-120B / MATH	250	1.94	7.74	1460953

I Long-Reference Sensitivity

This follow-up was specified before new generation. We reuse 50 main-study questions (MATH indices 50–99) but draw new Qwen3-8B Think references, at temperature 1, up to 16,384 tokens. Every reference is retained. For a reference of length T, the short view retains its first $\min(T, 4096)$ tokens; the long view retains all T tokens. We probe at 10% and 50% of each view and at a fixed position, token 409. Any position at or beyond the end is moved back to leave one reference token. Exact duplicate token positions share the same samples, preserving their dependence.

Each distinct position receives eight separately seeded samples for free, original-request, and close-reasoning continuation, all with an 8192-token output cap. This common cap differs from the original trace-dependent B_f. The design compares checkpoint definitions under a larger reference allowance; it is not a causal comparison of two independently executed reference-cap policies. The fixed-token condition holds prefix position fixed, whereas fractional checkpoints can move. Temperature and few-shot prompts are retained from the main study. This is a 16k sensitivity, not validation at the model card's 32k recommendation or on previously unused questions.

All 50 references end naturally (713–15,392 tokens; median 2372.5); 16 exceed the short view's 4096-token horizon. The fixed-token checkpoint (Fixed409 in tables) never needs this adjustment. Duplicate positions reduce 6,000 nominal continuations to 3,984 actual draws at 166 distinct pairs of question and prefix position. Including references, collection uses 4,034 calls and 7,724,910 generated tokens; the grader reports no timeouts.

The main revision effect remains large with full-reference checkpoints: 52.0 points at 10% and 13.0 at 50%. Free and original-request last answers have differences close to zero, with intervals spanning

zero at both positions; this is not evidence of exact equality. Closure markedly reduces the 10% gain, while its 50% contrast remains uncertain. Moving from the short to long view lowers the 50% gain by 6.5 [1.5, 12.8] points. These comparisons change checkpoint position within the same sampled reference, rather than identifying a causal effect of a reference-cap policy. Continuations still reach their 8192-token cap in some arms (Table 18); naturally ended references can still produce continuations that run out of tokens.

Table 16: **Three distinct questions under a longer reference allowance.** All entries are paired percentage-point contrasts with 95% whole-question bootstrap intervals ($n = 50$). Gain is original last minus first; Ordering is free-last minus original-last; Closure is original gain minus close-reasoning gain. Each interval describes one contrast and is not adjusted for multiple comparisons.

Checkpoint	Gain	Ordering	Closure
Short view, 10%	+53.8 [+41.2, +66.0]	-3.0 [-7.8, +0.5]	+45.2 [+33.8, +56.8]
Short view, 50%	+19.5 [+10.0, +30.0]	-1.0 [-3.2, +1.0]	+7.2 [-0.2, +15.5]
Long view, 10%	+52.0 [+39.2, +64.2]	+0.5 [-1.5, +2.5]	+41.2 [+28.7, +53.2]
Long view, 50%	+13.0 [+5.2, +22.0]	+0.2 [-1.0, +1.8]	+2.2 [-4.0, +9.2]
Fixed 409 tokens	+44.5 [+31.8, +57.2]	-1.8 [-6.5, +2.2]	+33.8 [+23.2, +44.5]

Table 17: **Moving from the short to the long reference view.** Paired long-minus-short differences in the three contrasts above, not differences between independently generated reference traces. Duplicate positions have exactly the same outputs. Fixed409 is a separate position control.

Fraction	ΔGain	ΔOrdering	ΔClosure
10%	-1.8 [-5.0, +1.0]	+3.5 [+0.0, +7.8]	-4.0 [-9.0, +0.2]
50%	-6.5 [-12.8, -1.5]	+1.2 [-1.0, +3.8]	-5.0 [-12.5, +0.8]

Table 18: **Complete per-arm horizon results.** Accuracy, answer coverage, and output-cap rates are percentages. First/last coverage is reported together; tokens are mean generated outputs. Missing answers count as incorrect. Every checkpoint and arm is retained, irrespective of direction.

Checkpoint	Arm	First	Last	Coverage F/L	Capped	Tokens
Short view, 10%	free	85.5	87.5	88.8/91.0	10.8	3184
Short view, 10%	original	36.8	90.5	100.0/99.8	8.2	2818
Short view, 10%	close	34.8	43.2	100.0/100.0	0.0	163
Short view, 50%	free	88.5	90.0	94.2/96.0	7.2	2308
Short view, 50%	original	71.5	91.0	99.8/100.0	5.8	1848
Short view, 50%	close	68.2	80.5	100.0/100.0	0.0	245
Long view, 10%	free	88.5	90.8	93.0/94.8	8.8	3130
Long view, 10%	original	38.2	90.2	99.8/100.0	8.0	2825
Long view, 10%	close	34.0	44.8	100.0/100.0	0.2	200
Long view, 50%	free	91.2	92.2	97.0/98.5	2.0	1954
Long view, 50%	original	79.0	92.0	100.0/100.0	2.2	1600
Long view, 50%	close	74.8	85.5	100.0/100.0	0.0	193
Fixed 409 tokens	free	85.8	87.5	88.8/91.2	10.2	3010
Fixed 409 tokens	original	44.8	89.2	99.8/99.8	8.2	2769
Fixed 409 tokens	close	39.0	49.8	100.0/100.0	0.0	208

Sampling logs, raw tokens, checkpoint mappings and source hashes are preserved. The frozen analysis resamples whole questions 20,000 times (seed 2026092101), keeping all arms and positions together. Reference termination, prefix-length distributions, and EOS/capped group summaries are retained in the analysis JSON. The study is offline: request times are records of the run, not controlled latency measurements. No human grading validation is claimed.

J Relation to Existing Readout and Stopping Protocols

The comparison below separates an answer request from a stopping policy. It records mechanisms described in the cited papers and the Dynasor-CoT project description; it does not infer undocumented

token limits or imply that all implementations use one answer parser. In particular, the 32/64/128 token examples in the Dynasor-CoT description are probing intervals, not verified output caps. For DEER, Algorithm 1 in the initial paper induces a trial answer before appending the end-of-thinking delimiter to the main path when the confidence threshold is met. Later versions or implementations may use different details.

Table 19: **Different protocols support different claims.** Published methods are not evaluated under our online service setup. "Boundary" means an explicit transition out of the model's designated reasoning section.

Work	Request, boundary, and decision	Relation to this audit
s1 (Muennighoff et al., 2025)	Budget forcing ends thinking at a maximum or extends it with "Wait" when the model tries to stop.	Establishes that generation control can change answers. Our audit partitions requested-output accuracy into immediate recovery and subsequent change.
NoThinking (Ma et al., 2025)	Prompts the model to bypass explicit thinking; compares token-limited and parallel generation.	Shows why an explicit reasoning boundary and total resource use matter. Our Qwen NoThink mode is not a reproduction of this method.
DEER (Yang et al., 2025b)	Induces trial answers at reasoning transitions, scores token confidence, then ends thinking when confident; includes branch-parallel acceleration.	Trial-answer confidence and terminal-answer accuracy are distinct readouts. A direct comparison must preserve its answer inducer, threshold, and branching policy.
Dynasor-CoT (Fu et al., 2025)	Appends answer-inducing text at checkpoints; checks consistency across checkpoints and post-answer uncertainty words.	Already inspects post-answer behavior before acceptance. Its across-checkpoint agreement differs from our simultaneous eight-vote rule.
Early answering (Lanham et al., 2023)	Shortens CoT and requests an answer to test how much reasoning is needed for the answer.	Motivates separating answer recovery from faithfulness; our question-only and endpoint controls address the interpretation of recovery.
This study	Free and requested outputs share sampling settings and allowed new tokens; score first/last new answers, coverage, changes, and delay.	Exposes the contribution of later answer changes under matched settings and reports the endpoint, channel, budget, and extraction dependence.

Our findings qualify the inference from a successful continuation to an immediately available answer; they do not establish that any cited method's reported results are incorrect. A direct reassessment would require its exact prompts, stopping rules, extraction, and resource accounting. For our own protocol, endpoint matching reverses the Qwen Think ordering under the tested cap, while channel closure changes the size of later gains. Both results must be reported with their cap and checkpoint dependence.

K Answer Transitions and Post-Answer Generation

Scope and data sources. This is a post-hoc analysis of the matched Qwen MATH and repaired replication studies, designed after their endpoint differences were known. It adds no model calls. We retain all eight free and eight forced outputs at both 10% and 50% own-prefix checkpoints: 14,400 unique continuations across five combinations of model, mode, and task. Scoring both strict and Label+ GPQA readouts produces 20,800 records, not additional independent samples. Every primary first/last label and correctness value matches the existing results. Raw-token and primary rendered endpoints agree for these Qwen own-prefix outputs. The analysis preserves all samples, including truncated outputs.

Illustrative trace. Figure 3 uses MATH index 140, forced sample 2 of 8, from the 10% Qwen Think condition. Its reference ends naturally at 1,496 tokens; the retained prefix is 149 tokens. The first box closes at new token 2 with answer 1, the first correct box at token 1,260 with answer 13, and the last box at token 2,226, again with 13. The output contains the verbatim sentence "Therefore, my

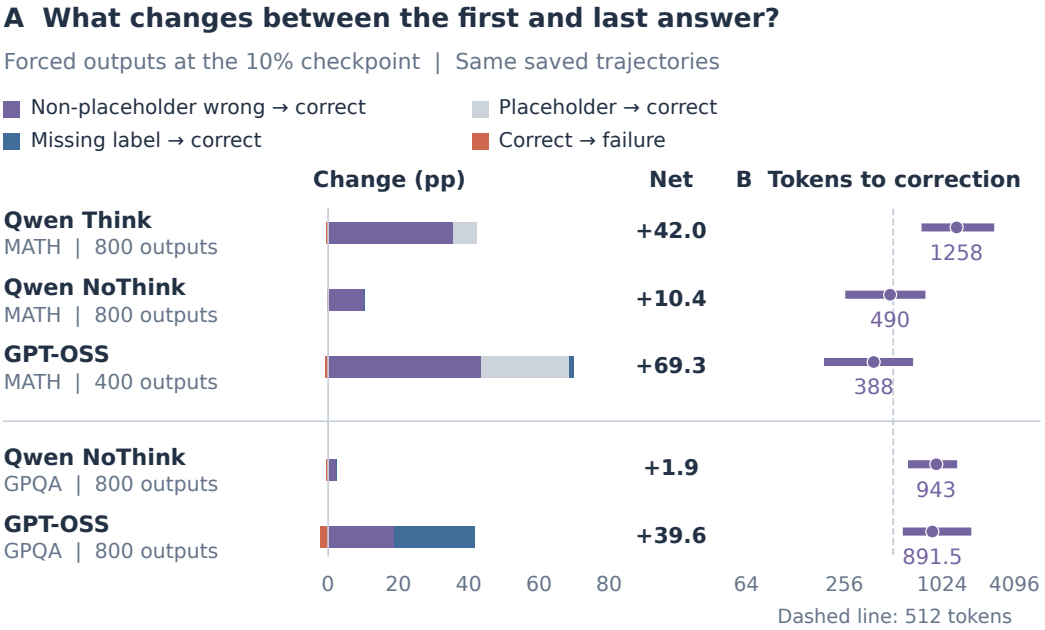

Figure 8: **Later generation changes the scored answer.** Forced outputs at the 10% checkpoint. (A) Bar segments show improvements and losses; *Net* is last-answer minus first-answer accuracy in percentage points. Gray separates placeholders such as `?` from other wrong answers. (B) Additional tokens to the first correct box among timed improvements: dots are medians and lines span the middle 50%, on a logarithmic axis. Qwen is 8B; GPT-OSS is 120B. The tables below give full counts and 50% results.

initial answer of 1 was wrong." It mentions 13 before the first correct box, so the 1,258-token delay measures boxed-answer availability, not the first appearance of the number or an internal decision. This example was selected post-hoc for simple numeric answers and a delay equal to the reported overall median; it is illustrative, not a frequency estimate. Raw-token offsets, the saved endpoint record, and source hashes accompany the figure artifact.

Accounting for changes. Equation 4 partitions last-minus-first accuracy into four correctness-changing transitions. The other five transitions have zero contribution. For a fixed problem set, free-last minus forced-first also equals free-last minus forced-last plus the forced last-minus-first change. Thus the observed 37.5-point Think gap at 10% decomposes into a -4.5-point last-answer contrast and a 42.0-point within-forced change. These are exact accounting identities, not a new theorem or a model of what causes what. Their value is making distinct observed events visible.

An extracted answer need not be a meaningful attempt. The unchanged MATH parser extracts placeholders such as `?`, `...`, and `something`; they belong to W, not M. Therefore, WC counts changes in scored answers, not exclusively repairs to an attempted mathematical solution. By contrast, strict GPQA WC requires replacing one recognized option letter with the correct letter. GPQA MC can reflect formatting, an answer without an explicit option label, or an incomplete first box; it is not evidence that the model previously lacked the answer. Label+ rescoring uses the same outputs and still finds both sources of gain.

For the main Think/MATH result at 10%, 286 of 339 gains (84.4%) begin with a non-placeholder wrong answer and 53 (15.6%) with an explicit placeholder; none lacks an extracted first label. The corresponding median delays are 1,209.5 and 1,352 tokens. Thus placeholder replacement is real but does not account for most of the reported change. This string-based audit avoids a subjective claim that every non-placeholder string is a meaningful solution attempt.

Blinded content sensitivity. We additionally gave GPT-OSS-120B, a model family different from the Qwen source, only the problem and first-answer string, hiding gold answers, later outputs,

Table 20: **Sources of first-to-last accuracy change in forced outputs.** Counts use all outputs in each row: W = extracted incorrect, M = missing label, C = correct. For example, WC means first wrong, last correct. Net change is in pp with paired 95% question-bootstrap intervals. GPQA rows without Label+ use strict letters. Qwen is 8B; GPT-OSS is 120B.

Cell / readout	Prefix	Total	WC	MC	CW	CM	Net [95% CI]
Qwen Think / MATH	10%	800	339	0	3	0	42.0 [33.5, 50.6]
Qwen Think / MATH	50%	800	85	0	2	0	10.4 [5.9, 15.5]
Qwen NoThink / MATH	10%	800	83	1	1	0	10.4 [7.1, 14.0]
Qwen NoThink / MATH	50%	800	117	0	2	0	14.4 [9.4, 19.9]
GPT-OSS / MATH	10%	400	275	5	3	0	69.3 [58.3, 79.5]
GPT-OSS / MATH	50%	400	119	1	1	0	29.8 [19.0, 41.3]
Qwen NoThink / GPQA	10%	800	17	2	2	2	1.9 [0.5, 3.5]
Qwen NoThink / GPQA	50%	800	12	11	3	5	1.9 [0.0, 4.0]
GPT-OSS / GPQA	10%	800	151	184	12	6	39.6 [31.9, 47.4]
GPT-OSS / GPQA	50%	800	63	35	14	8	9.5 [4.5, 15.0]
Qwen / GPQA Label+	10%	800	26	0	3	0	2.9 [1.1, 5.3]
Qwen / GPQA Label+	50%	800	18	15	4	0	3.6 [1.1, 6.5]
GPT-OSS / GPQA Label+	10%	800	152	191	12	5	40.8 [32.9, 48.5]
GPT-OSS / GPQA Label+	50%	800	63	36	14	8	9.6 [4.5, 15.1]

Table 21: **MATH gains split by the first generated answer.** The count columns are A = non-placeholder, P = placeholder, and M = missing. A non-placeholder is any extracted, nonempty first answer other than the three explicit strings `?`, `...`, and `something`; it does not assert that every string is a well-formed mathematical object. Missing means no first label was extracted. A/P delay columns report the 25th percentile / median / 75th percentile in additional generated tokens. They are descriptive conditional summaries, not confidence intervals.

Cell	Prefix	Gains	A	P	M	A delay	P delay
Qwen Think	10%	339	286	53	0	807.5/1209.5/1972.8	891/1352/2145
Qwen Think	50%	85	81	4	0	53/300/738	1091.8/1792.5/2572.5
Qwen NoThink	10%	84	83	0	1	274.5/490/763.5	–
Qwen NoThink	50%	117	117	0	0	93.8/161.5/242.8	–
GPT-OSS	10%	280	174	101	5	256.8/463.5/776.5	185/305/560
GPT-OSS	50%	120	102	17	1	249.5/356/513.8	74/240/329

correctness, repeated-pair counts, and the source model. Three judgments per pair covered all 96 distinct problem–answer pairs representing 286 non-placeholder gains.

At the pair level, majority vote labels 89/96 as concrete answer attempts. Because some pairs recur, this covers 273/286 events. The other pair-level labels are 1 malformed or non-answer, 2 placeholder or abstention, and 4 ambiguous. Requiring agreement from all three judgments labels 86/96 pairs, covering 254/286 events, as concrete attempts.

The judge produced 15 unparseable labels and hit its output cap 47 times across 288 judgments; both are retained. This post-hoc cross-model audit is neither a human audit nor a correctness regrade, and a concrete answer does not prove that a solution was complete.

Locating a revision in original tokens. We enumerate complete boxes with matching braces in the decoded saved token sequence. For each box, its closure position is the smallest original-token prefix containing the complete box, checked at the boundary and one token earlier. The suffix's injected opening is not charged as a generated token. For an output with an unsuccessful first answer and a correct last answer, we locate the earliest correctly graded box beginning after the initial box closes. The delay counts generated tokens between the closures of the first box and the first later correct box. Multiple closures in one token can give zero delay. A later box nested within the initial box is not a subsequent revision: one NoThink/MATH output at 50% has this pattern and is excluded from timing, while retaining its endpoint scores. An unclosed first box likewise gives no post-box delay.

Grading and uncertainty. Endpoint verdicts are retained, not re-selected. Previously ungraded intermediate MATH boxes use the same fixed mathematical grader with its two-second timeout:

Table 22: **Correct answers often require substantial post-answer generation.** "Gains" counts $WC + MC$; "Timed" requires a closed first box and a later, non-nested correct box. Delay quartiles list the 25th percentile / median / 75th percentile of additional tokens after the first box, not uncertainty intervals. The last column is the percentage of timed gains exceeding 512 tokens, with paired 95% bootstrap intervals. Undefined timing events remain in the endpoint analysis.

Cell / readout	Prefix	Gains	Timed	Delay quartiles	> 512 (%) [95% CI]
Qwen Think / MATH	10%	339	339	810/1258/2024.5	91.2 [85.3, 96.0]
Qwen Think / MATH	50%	85	85	60/317/903	37.6 [19.5, 60.0]
Qwen NoThink / MATH	10%	84	83	274.5/490/763.5	44.6 [26.9, 61.8]
Qwen NoThink / MATH	50%	117	116	93.8/161.5/242.8	4.3 [0.0, 10.2]
GPT-OSS / MATH	10%	280	275	203/388/641	37.5 [25.4, 50.2]
GPT-OSS / MATH	50%	120	119	221/341/497	22.7 [9.8, 38.5]
Qwen NoThink / GPQA	10%	19	19	668/943/1197	84.2 [65.0, 100.0]
Qwen NoThink / GPQA	50%	23	23	390/590/637.5	65.2 [33.3, 85.7]
GPT-OSS / GPQA	10%	335	332	620.8/891.5/1463	86.4 [79.2, 92.9]
GPT-OSS / GPQA	50%	98	97	616/859/1271	89.7 [83.6, 95.8]
Qwen / GPQA Label+	10%	26	26	599/832/1140	80.8 [66.7, 100.0]
Qwen / GPQA Label+	50%	33	33	472/589/736	69.7 [47.6, 84.8]
GPT-OSS / GPQA Label+	10%	343	340	623.3/898/1507	86.5 [79.3, 92.8]
GPT-OSS / GPQA Label+	50%	99	98	616/859.5/1312.3	89.8 [83.8, 95.9]

45 new pairs of candidate and gold answers in Qwen Think, 76 in NoThink, and 37 in GPT-OSS, with zero timeouts. GPQA compares extracted letters against the recorded key. We use 20,000 whole-question bootstrap draws (seed 2026091906), keeping arms and samples for each question together. Transition estimates average question-level proportions. Timing percentages are ratios of the total relevant events in each resample, not averages of question-wise ratios. A confidence interval is omitted if any bootstrap denominator is zero; a one-question group is descriptive only. Empirical 100% rates and bootstrap intervals that collapse to $[100\%, 100\%]$ are not population guarantees.

Complete reporting and interpretation. Artifacts retain all nine transitions, both free and forced outputs, both prefix fractions, both GPQA extraction rules, and naturally ended versus cap-truncated reference groups. Each sample retains its endpoints, boxed events, token positions, and timing exclusions. Tables here show forced outputs because they address the immediate-answer interpretation. Later checkpoints can have much shorter delays: for example, Think/MATH's median falls from 1,258 to 317 tokens. Timing is conditional on the saved output improving by its last answer; unsuccessful and unfinished outputs are not assigned a hypothetical correction time. It therefore does not estimate when all outputs would eventually succeed, nor does it compare newly sampled short-cap requests. Token distances do not identify internal computation or wall-clock latency.

L Pre-Specified Answer-Request Follow-ups

After observing the main endpoint result, but before making any calls for this follow-up, we fixed its problem set, prompts, seeds, sample count, cap rule, analysis, and success criterion. We sampled one new Qwen3-8B Think reference for each of MATH indices 0–49, separate from the main indices 50–149. At the 10% checkpoint, each shared prefix received eight distinct single-sample requests at temperature 1 in each of three arms: `Therefore, the final answer is \boxed{`, `The answer is \boxed{`, and `</think>` followed by `The final answer is \boxed{`. Each request used the same cap $\min(4096, \max(256, 2(T_{\text{obs}} - k_f)))$ as the main study. All 50 references and 1,200 continuations are retained; no answer or grade selected a reference or sample.

All three whole-question bootstrap intervals for last-minus-first accuracy exclude zero, satisfying the pre-specified directional criterion. The end-think request is nevertheless different in scale: its mean output is 189 tokens and 0.8% reach the cap, versus roughly 2,400 tokens and 31.5% for the two requests that leave the thinking region open. The follow-up therefore supports the existence of post-answer change across these requests, not the same effect size across requests or a claim about unobserved internal state.

Table 23: **Accuracy with three requests on separate problems.** First and Last are accuracy percentages; Change is in percentage points. Gain split counts non-placeholder / placeholder / missing events.

Answer request	First	Last	Change [95% CI]	Gain split
Original request	36.5	78.3	41.8 [30.3, 53.5]	151/15/1
Natural request	35.3	78.5	43.3 [30.8, 55.8]	161/14/0
End-think request	35.8	46.3	10.5 [5.8, 15.5]	44/0/0

Table 24: **Cost and event details for the answer-request follow-up.** Gain counts are non-placeholder / placeholder / missing first answers. Non-P and P abbreviate non-placeholder and placeholder. Delay columns give median additional tokens to the first later correct box; "–" means no event in that class. Capped and mean tokens describe all 400 continuations in an arm.

Answer request	Gain counts	Non-P delay	P delay	Capped (%)	Mean tokens
Original request	151/15/1	1158	1053	31.5	2391.8
Natural request	161/14/0	1280	1305.5	31.5	2430.4
End-think request	44/0/0	477.5	–	0.8	188.5

This follow-up confirms the main pattern on new questions, but broader project decisions predate it, so it is not a fully independent replication. It covers one model, benchmark, and checkpoint, and the three requests all open a LaTeX answer box. Other interaction styles may yield different trajectories.

Cross-model and checkpoint extension. Before these calls, we fixed two MATH model/question groups: the 100 saved Qwen3-8B Think references at indices 50–149 and the 50 saved GPT-OSS-120B references at indices 200–249. At both 10% and 50%, we newly sampled eight distinct single-sample continuations for the original request and for a model-specific request that closes the reasoning channel before opening the answer channel. Qwen uses `</think>`; GPT-OSS closes its named `analysis` channel and opens its named `final` channel. The arms share each prefix and use temperature 1 and the reference-dependent cap B_f from Section 3.2 for both models. The pre-specified primary quantity is Original last-minus-first change minus Close-reasoning last-minus-first change. Its point estimate is positive in all four cells, as predicted, although the Qwen 50% interval includes zero. All 4,800 continuations are retained.

We also test whether closure changes first-answer accuracy. We report paired Original-minus-Close differences in pp, with 95% whole-question bootstrap intervals. For Qwen, they are $+2.6$ $[-1.3, +6.8]$ at 10% and $+3.6$ $[-1.0, +8.4]$ at 50%. For GPT-OSS, they are -5.8 $[-11.3, -1.0]$ at 10% and $+2.0$ $[-5.5, +9.5]$ at 50%. Only GPT-OSS at 10% excludes zero: closure improves its first-answer accuracy by 5.8 $[1.0, 11.3]$ pp, much less than the 69.0-point difference in first-to-last gain.

Table 25: **Transitions in the cross-model closure extension.** Counts classify first-to-last correctness changes; W, M, and C mean wrong, missing, and correct. Cap is the percentage reaching the requested output limit.

Model	Prefix	Request	WC	MC	CW	CM	Cap (%)
Qwen3-8B	10%	Original	340	2	10	0	33.2
Qwen3-8B	10%	Close reasoning	101	0	0	0	0.5
Qwen3-8B	50%	Original	82	2	0	0	23.9
Qwen3-8B	50%	Close reasoning	77	0	0	0	0.6
GPT-OSS-120B	10%	Original	272	7	2	0	3.0
GPT-OSS-120B	10%	Close reasoning	1	0	0	0	0.0
GPT-OSS-120B	50%	Original	108	1	0	0	3.2
GPT-OSS-120B	50%	Close reasoning	2	0	0	0	0.0

The intervention closes an explicit text channel for reasoning; it does not prove that all further computation stops. The model-specific control strings are similar in purpose but differ in text, and the extension reuses existing questions and references instead of using another separate problem set.

Accuracy along saved token streams. Figure 4 is a post-hoc reconstruction of all 4,800 extension outputs, with no new model calls. At cutoff b, we decode the first $\min(b, L)$ generated tokens of

an output of observed length L, restore only the prompt-supplied `\boxed{`, and score the latest complete box with the frozen MATH grader. An unfinished later box leaves the previous complete answer in place; among nested complete boxes, the later opening position determines which is latest. Incorrect, blank, and missing answers score zero. We retain every output and carry its last observed answer after it ends or reaches its cap. Accuracy can fall when a later answer is wrong; this is not cumulative success at finding any correct answer.

We average the eight outputs within each problem and then average problems. The curves include every token count from 1 to 4,096. Pointwise 95% intervals use 20,000 paired question-bootstrap draws, sharing resamples across requests and checkpoints within a model (seed 2026092601). Shading interpolates intervals on a fixed logarithmic grid; the accuracy curves are unsmoothed steps. Literal decoding and parsing at 55,547 token cutoffs, including both sides of every box completion, agree with the event reconstruction. Every final endpoint matches the frozen extension record. A fixed token cutoff differs from first-answer accuracy because answers appear at different positions. These curves describe saved outputs; they do not estimate the effect of rerunning generation with a shorter cap or the cost of a stopping policy.

M Exploratory Pilot: Agreement after a Shared Start

Question and scope. The wrong unanimous answers in the online audit motivate a separate question: does agreement become more reliable when the same number of answers comes from more independently sampled starts? This pilot does not establish such a benefit. We report it to distinguish joint wrong agreement from error conditional on agreement, and to retain its negative and fragile results. The two cells use Qwen3-8B Think on 16 MATH-500 questions (indices 400–415) and GPT-OSS-120B with medium reasoning on 16 GPQA-Diamond questions sampled outside the earlier 100-question set (selection seed 2026091804). Each task is paired with only one model, so model and task effects cannot be separated; neither cell establishes generality.

Sampling and readout. For each question, we sample eight reasoning starts of up to 512 tokens. We take prefixes at token 128 and token 512 of each start and draw eight free continuations from each prefix. The 4096-token limit covers the entire path, including the prefix; 128 and 512 specify prefix lengths, not continuation budgets. We also draw eight fresh, uninterrupted paths with the same limit. Every actual draw uses a separate temperature-1 call with a distinct seed. If a start has already terminated, its answer is retained without further generation; it is not replaced. We read the last complete boxed answer in the *full path*, unlike the suffix-only endpoint in the readout study. Voting uses whitespace-normalized MATH strings or explicit GPQA option labels. Correctness is used only for offline scoring, not for choosing starts, answers, or thresholds. The original seeded-batch pilot is excluded; these results use the replacement sampling protocol described in Appendix H, not the duplicated batches.

Allocations and denominators. An allocation $r \times m$ takes r starts and m continuations per start, always producing eight answers ($rm = 8$). Thus 1×8 shares one start across eight continuations; 8×1 uses eight starts once each. The saved sample bank contains the collected starts and continuations. We average exactly over its uniformly selected subsets without replacement. This keeps the empirical distribution of any one selected answer identical across allocations but does not match their token costs. Let U mean that all eight selected answers have the same nonmissing label, and W that this agreed label is wrong. Write $\Pr(E)$ for the probability of event E under this subset selection. Table 26 separates wrong unanimity $\Pr(W \cap U)$, coverage $\Pr(U)$, and conditional error $\Pr(W \cap U)/\Pr(U)$. Conditional error asks what fraction of unanimous votes are wrong. We divide the pooled wrong-unanimity probability by pooled coverage, rather than average question-level ratios. Missing answers cannot form agreement and are not removed from the eight-answer denominator.

Primary contrast and uncertainty. The primary comparison is wrong unanimity for 1×8 minus 8×1 at depth 512, fixed before collection. Intervals use 10,000 paired whole-question bootstrap resamples (seed 2026091805). Each interval covers one listed contrast rather than all contrasts simultaneously; the intervals are exploratory and not adjusted for multiple comparisons. Table 27 also retains the 128-token sensitivity and conditional-error contrasts. All ratio resamples in this table have positive denominators. Zero observed MATH errors produce bootstrap intervals that collapse to a single value, not a certificate of zero population risk. Full threshold-setting intervals and the fixed 5/8–8/8 threshold grid are retained in the accompanying analysis artifacts.

Table 26: **All eight-answer allocations in the exploratory pilot.** Each cell has 16 questions; depth counts prefix tokens. Wrong and Error are $\Pr(W \cap U)$ and $\Pr(W \mid U)$ in percent; coverage is $\Pr(U)$. Costs are bank-based output-token estimates, not measured policy latency. Paired contrasts appear in Table 27.

Cell	Depth	$r \times m$	Wrong (%)	Coverage (%)	Error (%)	Output tokens
Qwen/MATH	128	1×8	0.00	57.81	0.00	21,009
		2×4	0.00	56.44	0.00	21,137
		4×2	0.00	56.40	0.00	21,393
		8×1	0.00	56.42	0.00	21,905
	512	1×8	0.00	58.59	0.00	18,322
		2×4	0.00	56.84	0.00	18,834
		4×2	0.00	56.35	0.00	19,858
		8×1	0.00	56.23	0.00	21,906
GPT-OSS/GPQA	128	1×8	9.38	39.84	23.53	15,862
		2×4	8.76	39.55	22.15	15,990
		4×2	8.49	39.27	21.61	16,246
		8×1	8.41	39.22	21.45	16,758
	512	1×8	10.16	40.62	25.00	13,347
		2×4	9.83	40.30	24.40	13,852
		4×2	8.68	39.24	22.11	14,863
		8×1	6.38	36.93	17.27	16,885

Table 27: **Shared minus independent starts: paired pilot contrasts.** Values are percentage-point differences with 95% bootstrap intervals. The symbol < 0 denotes a negative bound that rounds to zero at two decimals. Only the 512-token wrong-unanimity contrast is primary.

Cell	Depth	Wrong-unanimity difference	Conditional-error difference
Qwen/MATH	128	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	512	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
GPT-OSS/GPQA	128	+0.96 [$<$ 0, +2.76]	+2.08 [-0.08, +7.43]
	512	+3.78 [-0.38, +11.72]	+7.73 [-0.87, +28.55]

Fragility and threshold dependence. The GPQA point difference is sensitive to individual questions. A post-hoc diagnostic that reruns the analysis after omitting each question in turn changes the primary wrong-unanimity contrast from $+3.78$ to -0.14 points when question 174 is omitted; the primary analysis retains that question. The threshold grid, fixed during collection before complete-cell outcome analysis, also gives no uniform conditional-error ordering. At depth 512, requiring five of eight votes gives 26.58% conditional error for 1×8 and 29.19% for 8×1, opposite to the ordering at unanimity. Coverage is 61.72% and 63.47%, respectively. These are unequal-coverage, unequal-cost threshold settings, not evidence for selecting a deployment threshold.

Table 28: **Censoring and segmentation limit the pilot.** Cap-hit and missing-answer rates average over branched paths; terminal rates average over starts. Segmented minus fresh compares single-answer accuracy from a saved start-plus-continuation path with eight uninterrupted fresh paths per question, in percentage points with paired 95% intervals. An interval containing zero does not establish equivalence.

Cell	Depth	Cap hit (%)	Missing (%)	Terminal (%)	Segmented − fresh
Qwen/MATH	128	37.70	26.56	0.00	+0.68 [-6.45, +7.23]
	512	37.89	25.39	0.00	+1.95 [-5.96, +9.77]
GPT-OSS/GPQA	128	19.43	27.25	0.00	+1.37 [-3.32, +6.93]
	512	21.09	29.49	10.94	-0.10 [-5.27, +5.76]

What this pilot does not establish. The pilot has small problem sets, pairs each model with only one task, and frequently reaches the token limit (Table 28). It therefore leaves behavior under longer

generation budgets unresolved. More starts also cost more tokens. The pilot therefore supports neither a root-allocation method advantage nor a latency claim. Actual collection used 2304 calls and 5,373,733 output tokens for Qwen/MATH, and 2192 calls and 3,997,325 output tokens for GPT-OSS/GPQA. These bank-collection costs include fresh controls and both depths; costs of the excluded pilot and sampling diagnostics remain separate in the artifacts rather than being erased or attributed to a single allocation.

N Crossed Study of Shared Starts

Purpose and fixed problem set. The follow-up tests both models on both tasks with a longer token limit, addressing two limits of the pilot. Qwen3-8B Think and GPT-OSS-120B medium reasoning each receive the same 50 MATH and 50 GPQA questions. MATH indices are sampled from 416–499 (seed 2026091901); GPQA questions are sampled from the 82 remaining after the earlier 100-question set and 16-question pilot (seed 2026091902). These problem sets do not overlap with the controlled studies; the benchmarks are still public. Models, questions, checkpoints, metrics, and the complete analysis grid were fixed before outcome inspection. The study was designed after the inconclusive pilot, which remains reported separately.

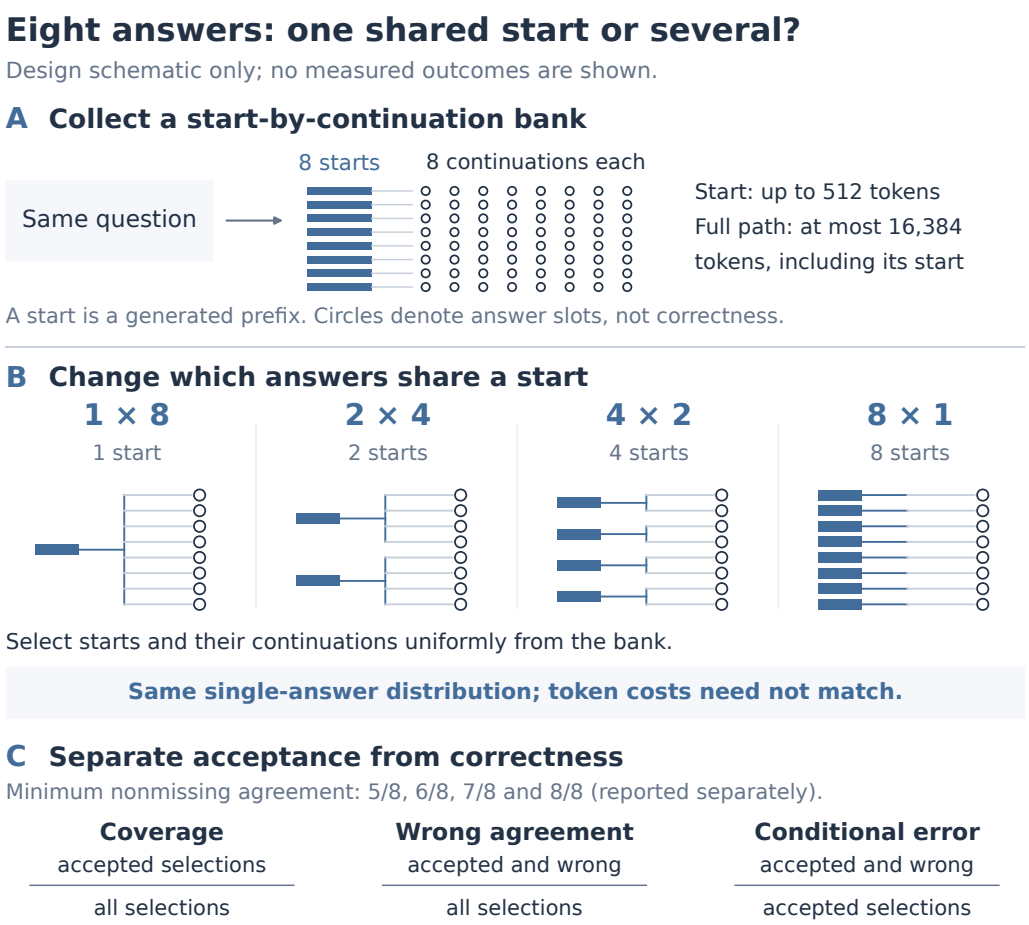

Figure 9: **Change shared history, not the number of answers.** An unfilled circle denotes a sample slot, not correctness. All four allocations select eight answers uniformly from the same collected starts and continuations, giving the same distribution for any one selected answer but different distributions over the eight answers together and different costs. If a start has already ended, its label is copied without a new model call. Correctness is used only after the observable vote has been selected.

Generation and readout. Each question receives eight independent 512-token starts, eight free continuations per nonterminal start, and eight fresh uninterrupted paths. The full-path limit is 16,384

tokens. All actual draws use separate calls at temperature 1 with distinct recorded seeds; naturally ended starts are retained, not resampled. We read the last complete box in the full path, including its prefix. This differs deliberately from newly emitted box readout in Section 3.2. MATH votes use whitespace-normalized strings; GPQA uses the explicit-label rule from Appendix F. Neither vote requires correctness. Missing answers cast no vote but keep their place in the denominator and their cost.

Selection and event definitions. An allocation $r \times m$ selects r of the eight starts and m of the eight continuations within each, uniformly without replacement, with $rm = 8$. We average over every valid selection in the saved sample bank, meaning the collected starts and continuations. Write $\Pr(E)$ for the probability of event E under this selection rule. For each threshold $t \in \{5, 6, 7, 8\}$, let E_t denote at least t identical nonmissing answers, and let W_t denote this event with an incorrect winning label. Coverage is $\Pr(E_t)$, wrong agreement is $\Pr(W_t)$, and conditional error is $\Pr(W_t)/\Pr(E_t)$. The ratio is undefined at zero coverage. These probabilities pool questions and selections, not question-specific error ratios. The primary contrast is $\Pr(W_8)$ for 1×8 minus 8×1 at the full horizon; no threshold is selected for deployment.

Uncertainty from the saved sample bank. We use 20,000 paired whole-question bootstrap draws (seed 2026091903). The equal-weight four-setting summary resamples within each task, keeping both models' outcomes for a question together: 100 question units, not 200 independent observations. Intervals cover individual contrasts, without multiple-comparison correction or guarantees for repeated online checks. Averaging all subsets removes selection randomness within the saved bank, not uncertainty from finite samples or the provider. Under independent starts and conditionally independent continuations, joint-event estimates are unbiased for their sampling-design probabilities; conditional-error ratios need not be.

Why more agreement need not imply lower error. For a fixed question and nonmissing label a, let C be a random start, $A_1, \ldots, A_8$ the eight selected continuation answers, and $p_a(C) = \Pr(A_i = a \mid C)$ for any selected answer i. Let $\mathbb{E}_C$ denote an average over random starts C. Under the independence assumptions above,

$$\Pr(A_1 = \cdots = A_8 = a) = \left(\mathbb{E}_C[p_a(C)^m]\right)^r, \qquad rm = 8. \tag{5}$$

The distribution of one answer, $\Pr(A_i = a) = \mathbb{E}_C[p_a(C)]$, is unchanged. Jensen's inequality, which here compares the average of a power with the power of an average, implies that sharing one start can increase unanimity for either correct or incorrect labels relative to eight starts. It does not order conditional error, which also depends on correct agreement. This standard identity is not a new theorem or an empirical effect-size estimate. Nor does it establish equivalence to uninterrupted generation.

N.1 Correctness Review and Full-Path Diagnostics

Primary and reviewed scores. The grader fixed before review remains the primary record. An AI-assisted review inspected all 621 model/question/label units, including positive, negative and missing labels and their box-payload variants. The review identified twelve false negatives: the grader marked these units wrong, but the review judged them correct. They comprise seven tuple or angle-bracket forms of $(-7, 16, 5)$ versus a column vector (MATH index 419), two reversed complete root lists $1, -2$ (index 456), two answers 15 with square-centimeter units implicit in the diagram (index 467), and $(x^9 - 1)/(x - 1)$ as the cancellable polynomial quotient (index 439). The latter treats the apparent division by zero at $x = 1$ as removable after algebraic cancellation; implicit units are an explicitly unit-tolerant convention. Wrong sign-reversed vectors, incomplete root lists, and different numerical answers remain wrong.

Reviewed verdicts apply to every occurrence across models, horizons, and reference methods: 749 endpoint occurrences change, with votes, coverage, costs, and missingness fixed. This is a post-hoc sensitivity, not human validation. The audit identified no MATH false positives. GPQA labels were checked against recorded keys, not independently re-solved; unrecognized formats remain missing. Equivalent mathematical expressions can still cast different string votes. Table 29 retains both grading conventions.

The correctness review weakens the apparent benefit of independent starts. Across equally weighted settings, the wrong-unanimity gap falls from 1.40 [0.22, 2.85] to 0.71 [−0.13, 1.76] points. The reviewed interval includes zero, and the conditional-error gap is also inconclusive. Thus this study

does not establish a general reliability advantage from allocating the same eight answers across more starts.

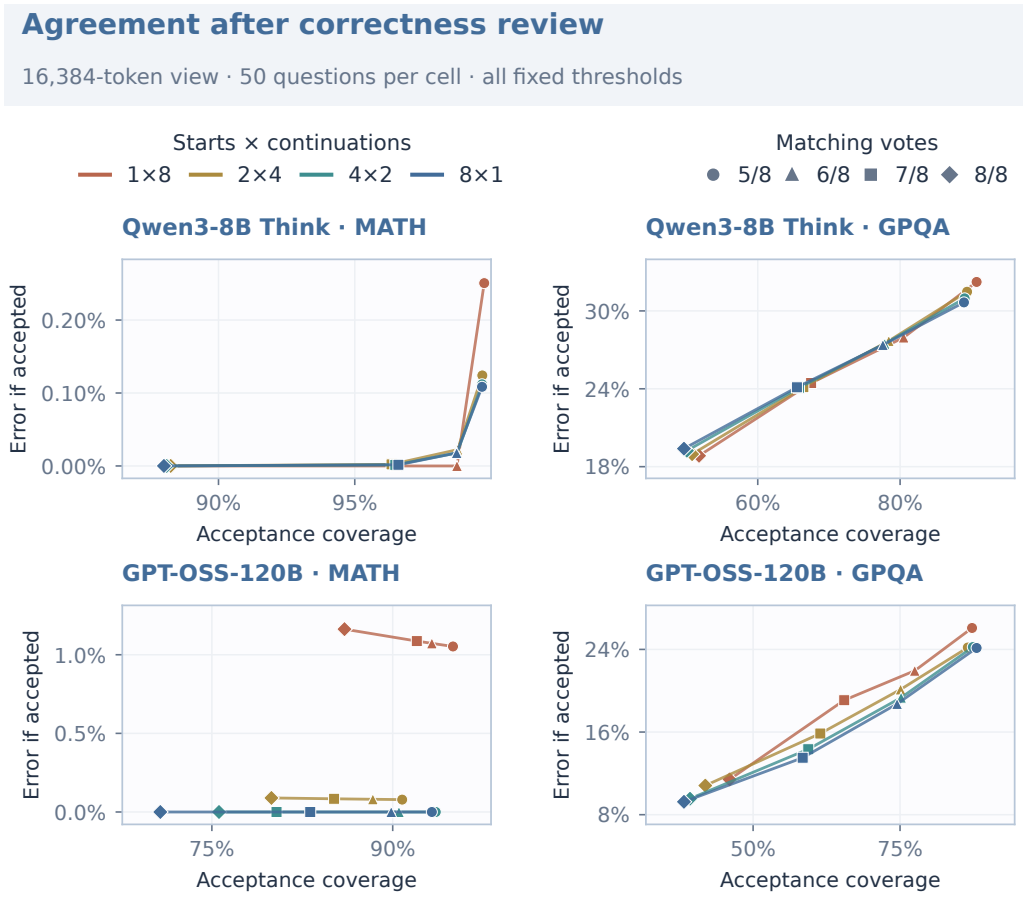

Figure 10: **Agreement thresholds change both coverage and conditional error.** All four allocations and all fixed thresholds at the 16,384-token horizon, after post-hoc correctness review. Curves connect thresholds, not matched-cost policies. Axes zoom separately; overlapping points are not artificially displaced. Terminal starts remain included. Every primary and reviewed interval, including the 4096/8192-token grids, is retained in the analysis artifacts.

Table 29: **A reliability contrast depends on correctness scoring.** 1×8 minus 8×1, in pp with paired 95% intervals; 50 questions per cell, 100 paired question units overall. Primary uses the unchanged grader; reviewed changes correctness only. Coverage and costs remain unequal. Values rounding to zero are not safety bounds.

	Wrong-unanimity gap		Conditional-error gap
Cell	**Primary**	**Reviewed**	**Reviewed**
Qwen/MATH	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Qwen/GPQA	0.13 [-0.43, 0.81]	0.13 [-0.43, 0.81]	-0.55 [-2.09, 0.85]
GPT-OSS/MATH	3.74 [0.74, 8.24]	1.00 [0.00, 2.25]	1.16 [0.00, 2.69]
GPT-OSS/GPQA	1.71 [-1.51, 5.92]	1.71 [-1.51, 5.92]	2.16 [-4.86, 10.50]
Equal-cell overall	1.40 [0.22, 2.85]	0.71 [-0.13, 1.76]	0.55 [-0.65, 1.97]

Table 30: **All full-horizon unanimity allocations after correctness review.** Coverage, wrong unanimity and conditional error are percentages; brackets give separate 95% question-bootstrap intervals for each row. Output tokens count each selected start once and every selected continuation. Costs and coverage are not matched. Near-zero rounded errors are not population-risk bounds.

Cell	$r \times m$	Coverage	Wrong	Error	Output tokens
Qwen/MATH	1×8	88.2	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	30,517
	2×4	88.2	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	31,029
	4×2	88.1	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	32,053
	8×1	88.0	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	34,101
Qwen/GPQA	1×8	51.7	9.8 [3.8, 17.0]	18.8 [7.4, 33.1]	66,437
	2×4	50.8	9.6 [3.7, 17.0]	19.0 [7.3, 33.4]	66,949
	4×2	50.0	9.6 [3.6, 17.0]	19.2 [7.4, 33.8]	67,973
	8×1	49.6	9.6 [3.6, 17.0]	19.4 [7.4, 34.1]	70,021
GPT-OSS/MATH	1×8	86.0	1.0 [0.0, 2.2]	1.2 [0.0, 2.7]	3,858
	2×4	79.9	0.1 [0.0, 0.2]	0.1 [0.0, 0.3]	4,272
	4×2	75.6	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	5,099
	8×1	70.7	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	6,753
GPT-OSS/GPQA	1×8	46.0	5.2 [1.2, 10.2]	11.4 [2.8, 22.5]	14,325
	2×4	41.9	4.5 [1.5, 8.4]	10.8 [3.5, 21.0]	14,834
	4×2	39.3	3.7 [1.0, 7.4]	9.5 [2.4, 19.8]	15,853
	8×1	38.2	3.5 [0.7, 7.2]	9.2 [1.9, 19.7]	17,891

Terminal starts and question influence. Terminal-start rates are 0% in both Qwen cells, 51% in GPT-OSS/MATH, and 3.75% in GPT-OSS/GPQA. Corresponding prefix box-presence rates are 0.25%, 0%, 66% and 10.75%. A complete box does not necessarily mean the trace ended. Splitting wrong unanimity by whether a start has ended retains the original denominators. This describes the saved starts rather than isolating the effect of ending a start. In GPT-OSS/MATH, reviewed wrong unanimity for 1×8 is 1.00%: 0.75 points involve an ended start and 0.25 involve only nonterminal starts. We retain all of them. All per-question contrasts and leave-one-question-out diagnostics are retained. For example, the GPT-OSS/GPQA primary mean is 1.71 points, but falls to 0.22 when question 75 is omitted; this diagnostic does not authorize deleting that question.

Table 31: **Short readouts remain substantially censored.** Missing and censored percentages average over branched paths; an unfinished path can already contain an answer. "Segmented" means a saved start followed by a continuation; "fresh" means an uninterrupted path. Accuracy contrasts are reviewed, segmented minus fresh, in pp with paired 95% intervals. Short horizons truncate saved long paths, not independently sampled shorter-cap requests.

Cell	Horizon	Missing (%)	Censored (%)	Segmented − fresh
Qwen/MATH	4,096	26.8	40.9	0.6 [-1.8, 3.5]
	8,192	7.5	11.1	1.1 [-1.5, 3.8]
	16,384	0.3	0.7	-0.1 [-0.7, 0.7]
Qwen/GPQA	4,096	85.2	87.3	-0.1 [-2.4, 2.1]
	8,192	49.4	54.3	0.4 [-3.2, 4.0]
	16,384	2.3	2.9	1.0 [-1.5, 3.5]
GPT-OSS/MATH	4,096	0.5	1.0	0.5 [-1.0, 2.4]
	8,192	0.0	0.0	0.0 [-1.0, 1.1]
	16,384	0.0	0.0	0.0 [-1.0, 1.1]
GPT-OSS/GPQA	4,096	14.5	11.5	-0.5 [-4.3, 3.2]
	8,192	5.8	1.2	-0.8 [-4.8, 3.2]
	16,384	4.6	0.0	-1.2 [-5.4, 2.9]

N.2 Independent Voting and Prefix Consistency References

Algorithms and implementation differences. Self-consistency (SC) selects the most frequent nonmissing label among fresh, uninterrupted paths. Prefix Consistency (PC) (Iwase et al., 2026) uses one regeneration per fresh path. After removing the leading reasoning channel marker, if present, we re-tokenize its text. For original token count L, remove $d = \lfloor L/4 \rfloor$ tokens and retain $L - d$, where the floor brackets round $L/4$ down to the nearest integer. The new-token cap is $\max(2d, 1638)$, while reserving 256 tokens to avoid exceeding a common 32,768-token context limit. No prompt needed shortening to fit this limit. Original full traces have their full generation cost counted before truncation; PC regenerated paths can exceed 16,384 tokens. The formulas match the pinned implementation, but our initial cap is 16,384 rather than 100,000 tokens. Prompts, provider token accounting, model availability, and missing-answer retention also differ. These are reference implementations, not a reproduction of published benchmark performance.

Within each original/regenerated pair, a label's frequency is 1 if both answers match it and $1/2$ if only one does. PC raises this frequency to power $p \in \{1, 2, 3\}$ for each nonmissing label. We sum these weights across pairs, retain missing members in the pair denominator, and average uniform tie-breaking over the highest-weight labels. Gold never chooses a winner. Both SC and PC are evaluated on all uniform subsets of $N \in \{1, 2, 4, 8\}$ original paths or pairs. Expected subset costs are not executions under a common hard resource budget.

Table 32: **Eight-original-path references, with their full costs.** Accuracy (%) before/after correctness review, with separate 95% intervals per row. PC includes eight extra regenerations. Powers 1–3 yield identical eight-pair accuracies, not independent replications. Input costs assume no cache reuse; compute is unequal.

Cell	Method	Primary	Reviewed [95% CI]	Output	Input
Qwen/MATH	SC	98.0	100.0 [100.0, 100.0]	34,975	1,478
	PC	98.0	100.0 [100.0, 100.0]	44,932	29,185
Qwen/GPQA	SC	65.0	65.0 [52.0, 78.0]	68,861	2,117
	PC	66.0	66.0 [52.0, 78.0]	87,650	55,878
GPT-OSS/MATH	SC	95.0	100.0 [100.0, 100.0]	6,904	1,921
	PC	96.0	100.0 [100.0, 100.0]	8,799	9,005
GPT-OSS/GPQA	SC	74.0	74.0 [62.0, 85.0]	18,216	2,532
	PC	73.0	73.0 [61.0, 84.0]	24,174	18,712

Why some reference curves coincide. Let c_a count pairs containing label a twice and d_a pairs containing it once. PC assigns $S_p(a) = c_a + 2^{-p} d_a$. When $N \leq 2^p$, the winners rank first by c_a, then by d_a: a one-pair lead exceeds the contribution of the at most $N - 1$ opposing singleton occurrences. Thus all powers agree for $N = 1, 2$, and powers 2 and 3 agree for $N = 4$. At power 1, PC selects the most frequent label (a plurality vote) on the paired answers, not on $2N$ independent paths. These identities explain overlaps; they do not imply accuracy or compute superiority.

Collection and reproducibility. The follow-up used 15,848 distinct-seed single-response calls, 54,332,323 output tokens and 14,510,627 input tokens counted without cache reuse. Concurrency was at most 32 requests per job and 8 in PC stages. Public list prices give an upper bound of $38.29, not an invoice or per-policy cost. Request timings document the run, not latency performance. All 50 questions per cell completed without application retries or outcome-based replacements. An offline parsing fix handles a Unicode line separator in one valid JSON data string; raw files, generation, graders and analysis formulas remain unchanged. Raw/result file fingerprints, primary analyses, the separate post-hoc review and all fixed grids are retained. Tokenizer files were not separately hashed; request audits cannot certify service-side random-number independence or exact future replay.

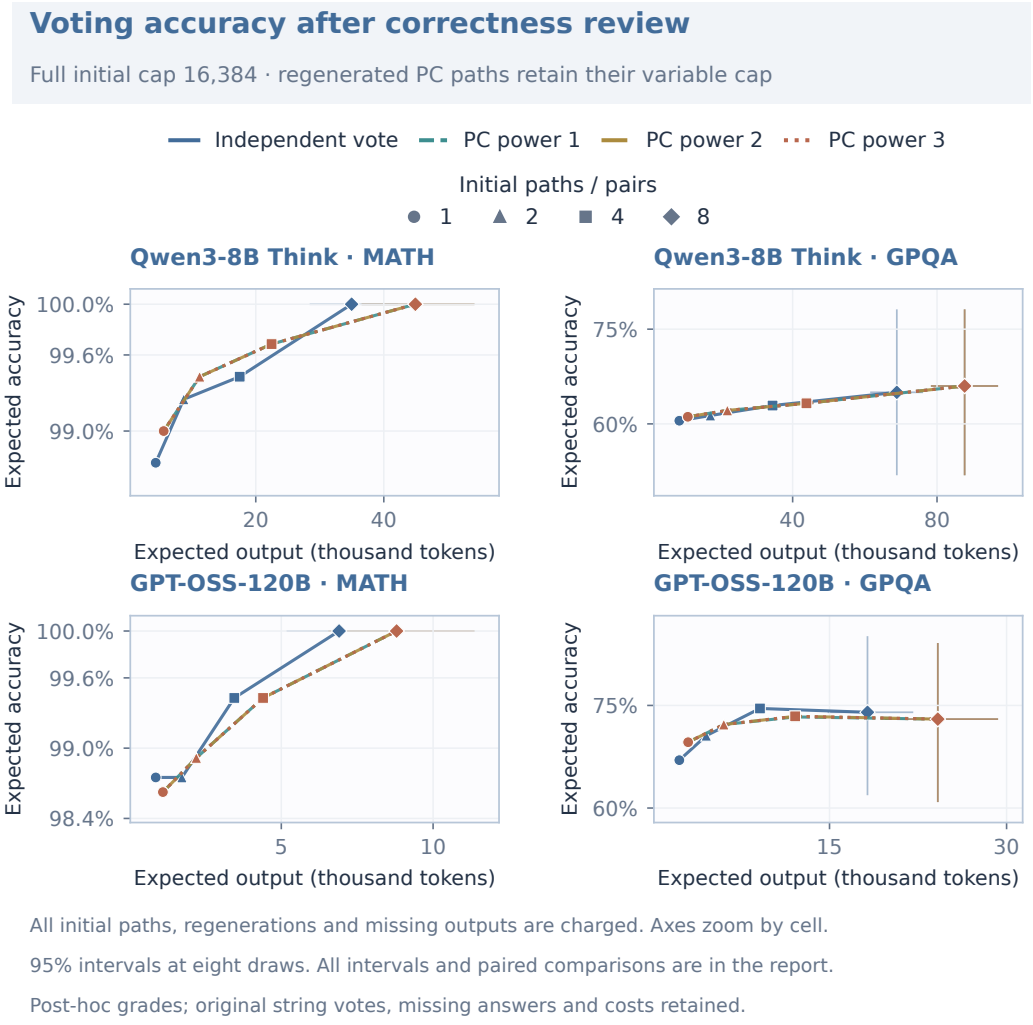

Figure 11: **Accuracy and full generation cost for the fixed reference grid.** All subset sizes and PC powers after correctness review; separate 95% intervals at eight original paths. Missing outputs retain their cost. PC overlaps reflect identical outcomes or voting identities, not omitted methods. Lines connect measured points without establishing continuous costs, matched-resource advantages, or latency.