REASONING, ANSWER DELIVERY, AND BUDGETED ACCURACY

Anonymous authors
Paper under double-blind review

ABSTRACT

A late answer cue can raise a reasoning model's score, but its interpretation depends on the comparison. We compare direct continuation D with a four-token cue L on the same prompt and sampled prefix, under one output cap that includes forced tokens. Correct delivery requires a natural end of sequence (EOS), a complete final boxed answer and acceptance by the fixed scorer, which allows numeric tolerance on MATH. On 200 unused MATH questions, preregistered L raises delivery by 7.33 percentage points, 95% interval [5.00, 10.00], for Qwen3.5-9B at 2048 tokens with 45 rescues and one harm, by 6.83 for Ministral-3-8B and by 4.50 for Qwen3.5-9B at 4096 tokens. Registered GPQA-Diamond transfers give +20.20 [16.50, 24.07] with Qwen and +4.55 [2.69, 6.57] with Ministral without harms, and L generates 1–4% fewer tokens. Only runs active at the cue can change, and across nine settings 23–49% of them convert net of harms. L exceeds a budget-charged variant of the closest protocol's conclusion message by 3.50 points [1.50, 5.67], largely through greater answer room, and a numeric-budget system prompt by 9.50 [6.17, 13.00]. The comparison changes reserve selection: the earliest cue looks best against the stopped stream in three MATH settings, but against continuation to the same cap a 128-token reserve beats it by 15–21 of 600 runs at 2048 tokens. Prospectively, at a 1024-token cap, a 128-token reserve beats a 512-token one by 14.33 [10.67, 18.33] and 10.33 [6.50, 14.33] points for the two models, while matched no-cue continuations deliver within 0.67 points of D. Continuing D also beats restarting thinking at 1024 tokens by 8.25 points at the same cap.

1 INTRODUCTION

A reasoning trace can contain a correct value yet fail to deliver an answer before its output limit. Conversely, forcing a final answer can improve a short-prefix score while losing to a longer uninterrupted continuation at the same total budget. These two distinctions matter whenever a service must choose between direct answering, continued thinking and an answer instruction.

Prior work compares direct answering, reserved answer stages and forced thinking termination at low budgets (Ma et al., 2025; Muennighoff et al., 2025; Xu et al., 2026), and inserts a conclusion message shortly before a strict output limit, compared with generation run to that limit (Sun et al., 2025). Three controls are easily conflated here: the same *nominal* budget, a per-call limit after which answer tokens may still be appended; the same *total* output cap with every forced token counted inside it; and the same *prompt and sampled trajectory* up to the intervention, compared with that trajectory continued. Our late control L versus direct continuation D imposes all three. We ask what such a paired, fully charged execution identifies, which runs it changes, and whether it changes conclusions that a looser comparison would draw.

We run a five-policy development matrix on 128 MATH questions, a fixed-policy rerun and a preregistered rerun on 200 unused questions each, the latter with a second model family, a looser cap and a budget prompt, and registered studies of cue wording and reserve, GPQA-Diamond (Rein et al., 2024) transfers, direct comparisons with the closest prior alternatives and a prospective reserve test; Table 1 lists them. With Qwen3.5-9B and a 2048-token total output allowance, L inserts on a new line the four-token cue "Final answer:" at token 1916 of the direct stream. Our contributions are:

1. **A localized, bounded and replicated delivery gain.** Only runs still active at the cue can differ, and their share bounds the effect by Proposition 2. In the preregistered rerun L gains $+7.33$ points

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.

[5.00, 10.00] over D with 45 rescues and one harm, $+6.83$ for Ministral-3-8B and $+4.50$ for Qwen3.5-9B at a 4096-token cap, and registered transfers to GPQA-Diamond give $+20.20$ [16.50, 24.07] with Qwen and $+4.55$ [2.69, 6.57] with Ministral, both without harms; the development set at $+7.81$ and the fixed-policy rerun at $+7.41$ agree. In all nine settings of Table 5 the gain lies in active runs, 23–49% of which are converted net of harms.

2. **Direct comparisons under one accounting.** On the same streams and cap, the registered cue exceeds our budget-charged variant of the concluding message of Sun et al. (2025) by $+3.50$ points [1.50, 5.67] with Qwen at 2048 tokens, mostly because that position leaves 11 answer tokens, and exceeds a system prompt that states the numeric budget by $+9.50$ [6.17, 13.00]. At the same position the field cue also exceeds each of its components alone: a box opener, a deadline and a neutral field label.
3. **The comparison decides the budget conclusion.** Against the stream stopped at the cue, a 512-token answer reserve has the largest gain in all three MATH settings; against the stream continued to the same cap, a 128-token reserve gains more at 2048 tokens, 42 and 43 net runs of 600 against 21 and 28 at 512, and matched no-cue continuations at every reserve deliver within 0.67 points of D. A reserve-exchange identity, Proposition 3, attributes the difference mainly to the conversion of the runs both cues reach. A registered prospective test in operating conditions without earlier reserve outcomes confirms the short reserve at a 1024-token cap, where 128 tokens beat 512 by $+14.33$ and $+10.33$ points for the two models after Holm correction, and the 64-token reserve that three of four frozen predictors chose delivers most in all three new conditions.
4. **The identified effect is correct delivery, and continuing beats restarting.** Of 138 capped direct outputs on unused MATH questions, 137 end without a complete final boxed field, as do 357 of 359 on GPQA-Diamond. At the same cap on unused questions, continuing the direct stream beats discarding its prefix at 1024 tokens and restarting thinking by 8.25 points, because the restart forfeits 62 correct deliveries of streams that would have finished.

2 Related Work

Answer forcing at a budget. Ma et al. (2025) append a final-answer prefix once a generation exhausts its token limit, so the answer tokens fall outside that nominal limit, and compare thinking with bypassing it at matched budgets. s1's budget forcing (Muennighoff et al., 2025) and the Qwen3 thinking budget (Yang et al., 2025) end a thinking trace and ask for the solution; token-level triggers can also steer hybrid models between modes (Yang et al., 2026b). Closest to our setting, Sun et al. (2025) tell the model in its prompt that it may be interrupted, insert a concluding message and allow 25 answer-generation tokens, and compare it with generation run directly to that limit, extracting the answer from the output, over 30 models. Their baseline, like our D, runs to the same limit. The protocols differ in three controls: the prompt notice changes the generation before the cue, whereas our L and D share the prompt and every sampled token up to the cue, which shows which runs a forced answer can change, namely those active at the cue by Proposition 2; we count forced tokens inside one total cap shared with restart alternatives; and we score delivery, meaning EOS and a final field, separately from extraction. Section 6 tests budget-charged variants of their message on our streams.

Interruption, leakage and incomplete chains. Wu et al. (2026) interrupt reasoning models with end-thinking or force-answer text and document reasoning leaking into the answer and answers followed by continued reasoning without EOS; hybrid models also mix thinking into their no-think mode (Wang et al., 2025), and partial chains can hurt more than no reasoning (Su et al., 2026). We make EOS and a single complete final field part of the endpoint, so leaked or continued reasoning after an answer counts as a delivery failure rather than a success.

When is the answer formed? Early answering truncates a chain of thought and elicits an answer to test whether the reasoning is post hoc (Lanham et al., 2023), and forced completion at partial prefixes shows that answers are often fixed well before the trace ends (Datta et al., 2026). Early-exit methods probe intermediate answers and stop when they stabilize or become confident (Fu et al., 2025; Yang et al., 2026a; Liu & Wang, 2025; Hosseini et al., 2026); CoDE-Stop also reuses base trajectories and counts the cost of probing, and internal activations can partly predict non-convergence (Oladri et al., 2026). These methods decide *when* to stop from answer signals; we fix one cue position, pair it

Table 1: Studies, question sets and roles. Questions come from the MATH test split (Hendrycks et al., 2021), except the registered transfers to GPQA-Diamond (Rein et al., 2024); every study uses three fixed sampling streams per question, a total output cap including forced tokens, and the question's three-stream mean as the unit of inference. Arms: D direct, L late answer cue, T thinking, F forced thinking close, C restart thinking; P and Pn add budget instructions to D's system prompt. The model is Qwen3.5-9B and the cap 2048 tokens, except for the second rerun's Ministral-3-8B and 4096-token experiments, its registered 1024-token secondaries of Appendix D, the Ministral arms of the later studies and the 1024-token continuations of the fourth study in Appendix H. Appendix Table 34 gives the verbatim prompts and forced texts.

Study	Questions	Arms	Endpoint	Role
Development, §5	128, exposed	D L T F C	delivery	development
First rerun, §6	200, unused	D L T F C	delivery	fixed policies
Second rerun, §6	200 further, unused	D L P; D L at 4096	delivery	preregistered
Follow-up, §6	second rerun's	D streams re-continued	delivery	registered
Transfer, §6	198 GPQA-Diamond	D L	delivery, letter	registered
Third study, §6	second rerun's; GPQA	D streams re-continued; Pn	delivery	registered
Fourth study, §6	GPQA; second rerun's	D L; D re-continued	delivery	registered

with the same sampled continuation under one total cap, and ask which runs change and on which endpoint.

Trained budget control, routing and cost-aware evaluation. Elastic Reasoning (Xu et al., 2026) and AnytimeReasoner (Qi et al., 2025) train models under truncated thinking budgets; Adapt-Think (Zhang et al., 2025) learns when to think, DART (Lee et al., 2026) routes budgets from agreement between cheap drafts, and HRBench (Ning et al., 2026) compares prompt-based, routing and speculative mode-switch strategies under one protocol. Our interventions require no training; restart policy C is one fixed route. Test-time compute helps in a problem- and allocation-dependent way (Snell et al., 2025), and longer reasoning can be wasteful or harmful (Chen et al., 2025; Gema et al., 2025; Zhou et al., 2026). Dong et al. (2026) show that a stopping rule saves or costs tokens depending on cache reuse; we charge restarts for discarded generation and a second prefill. Prompt-level budget statements (Han et al., 2025) change the whole generation, whereas L leaves every token before the cue unchanged; we compare both. Concurrently, Carlon et al. (2026) pair seed-matched runs to evaluate concise or early-answer instructions, and Guedes de Souza & Panisson (2026) show that model rankings reverse across budgets.

3 Answer Contracts and Paired Policy Evaluation

Let B be a total generated-token allowance. Direct answering D and thinking T receive the same question, but their native chat templates differ. F retains a thinking prefix and adds a close-plus-answer instruction; the late answer cue L retains a direct prefix and adds only the answer instruction. Gold answers are used by the scorer, never by the policy.

Three outcomes. A permissive endpoint extracts an answer from available text. Our primary endpoint E1 requires a natural end-of-sequence token, EOS, within the cap, a complete last balanced boxed answer after any thinking close, no substantive trailing text, and acceptance by the fixed scorer. MATH scoring compares normalized strings and symbolic expressions, also accepting numeric values p, g when $|p - g| \leq 10^{-4} \max(1, |g|)$. GPQA requires the gold option letter. Thus "correct delivery" means operational scorer acceptance, not exact semantic equivalence; the audits of Appendix B.3 assess that distinction.

Three costs. We report the maximum total output allowance, realized generated tokens, and measured execution time separately. Forced instructions count toward B. A restart charges its discarded generation and second prefill. Arms of one batch run consecutively on the same model instance, so batch-policy wall compares arms within the tested workload; it is not per-request service latency.

4 What the Accounting Identities Imply

For a fixed model, question distribution and decoding procedure, let L denote the unconstrained natural output length in thinking mode, with cumulative distribution function $F_L(b)$. Let $\alpha_c(b)$ and $\alpha_t(b)$ be correctness conditional on $L \le b$ and $L > b$, under a specified scorer. These definitions concern total natural output length; the last thinking-close position is a separate observable.

Proposition 1. Accuracy accounting.

$$\mathrm{Acc}_{\mathrm{T}}(b) = F_L(b)\alpha_c(b) + (1 - F_L(b))\alpha_t(b). \tag{1}$$

The proof is total probability; conditioning on completion changes the question population.

Shared prefix versus equal total budget. Suppose naturally completed outputs are returned unchanged and an alternative answer procedure has conditional accuracy $\alpha_e(b_r, b_a)$ on unfinished prefixes. Its gain over stopping the original procedure at b_r is

$$\Delta_{\mathrm{prefix}} = (1 - F_L(b_r))[\alpha_e(b_r, b_a) - \alpha_t(b_r)]. \tag{2}$$

It is nonnegative when the bracket is nonnegative. At the same *total* allowance $b = b_r + b_a$, however, subtraction gives

$$\begin{aligned}\Delta_{\mathrm{total}} = {} & (1 - F_L(b_r))[\alpha_e(b_r, b_a) - \alpha_t(b)] \\ & - [F_L(b) - F_L(b_r)][\alpha_c(b) - \alpha_t(b)] \\ & + F_L(b_r)[\alpha_c(b_r) - \alpha_c(b)].\end{aligned} \tag{3}$$

The completion and composition corrections cannot generally be discarded. For

$$F_L(b_r) = .1, \quad F_L(b) = .9, \quad \alpha_c = 1, \quad \alpha_t = 0, \quad \alpha_e = .5,$$

the shared-prefix gain is .45, yet split accuracy is .55 and continued thinking achieves .90: the equal-total difference is $-.35$. The live study therefore measures continued D directly.

Localization under prefix identity. The live control is paired at the token level, which gives a sharper identity than Eq. (3).

Proposition 2. Localization. *Let L insert its cue at position $t < B$ into the same sampled stream as D, and let A be the event that D has not emitted EOS by t. If both policies produce identical tokens up to t, then*

$$\mathrm{Acc}_{\mathrm{L}} - \mathrm{Acc}_{\mathrm{D}} = \Pr(A)\,\big[\Pr(\textit{L correct} \mid A) - \Pr(\textit{D correct} \mid A)\big], \tag{4}$$

and, under an EOS-gated endpoint,

$$\begin{aligned}\mathrm{Acc}_{\mathrm{L}} - \mathrm{Acc}_{\mathrm{D}} &\ge -\Pr(A,\ \textit{D ends correctly in } (t, B]), \\ \mathrm{Acc}_{\mathrm{L}} - \mathrm{Acc}_{\mathrm{D}} &\le \Pr(A) - \Pr(A,\ \textit{D correct}).\end{aligned}$$

Proof. Outside A, both policies return the same completed output. Inside A, D is correct only if it ends correctly in $(t, B]$, because a capped run delivers nothing under the contract. Total probability gives the rest. □

$\Pr(A)$ is measurable from D alone, so the active share bounds the effect and forces it to vanish where direct streams rarely approach the cap; the sign and size within A remain empirical. On the development questions $\Pr(A) = 66/384$ and only 7 D runs end in $(1916, 2048]$; on the unused questions the counts are 143/600 and 5. The observed changes, 30 and 45 rescues with 0 and 1 harm, lie inside these bounds, as they must.

Choosing the answer reserve.

Proposition 3. Reserve exchange. *Let L and L′ insert the same cue into the same sampled stream at positions $t' < t$, corresponding to reserves $r' > r$, and with active events $A' \supseteq A$. Under prefix identity,*

$$\begin{aligned}\mathrm{Acc}_{\mathrm{L}'} - \mathrm{Acc}_{\mathrm{L}} = {} & \Pr(A' \setminus A)\ \Pr(\textit{L' correct} \mid A' \setminus A) \\ & - \Pr(A' \setminus A)\ \Pr(\textit{D correct} \mid A' \setminus A) \\ & + \Pr(A)\big[\Pr(\textit{L' correct} \mid A) - \Pr(\textit{L correct} \mid A)\big].\end{aligned}$$

Table 2: Complete live matrix: 128 exposed questions, three fixed sampling streams, five arms; total output cap 2048 including instructions. Prefill is the mean number of actual prompt tokens; output includes all forced tokens. Correct delivery is the boxed-final-answer plus EOS contract. Wall is summed batch-policy execution time, including prefill and intervention overhead, not request latency.

Policy	Correct/384	Correct by seed/128	EOS/384	Mean output	Prefill	Wall, s
D: direct	308	100,103,105	325	934.63	98.77	4614.92
L: late answer cue	338	114,110,114	379	916.73	98.77	4414.47
T: thinking	76	27,26,23	76	1911.39	96.77	4972.54
F: forced thinking close	254	85,83,86	375	1070.44	96.77	3217.75
C: restart thinking	259	90,84,85	373	1015.83	140.92	4587.15

The first two lines are the price of an earlier cue: runs that D would finish between the two positions are handed to the cue, and those D delivers correctly can only be kept or lost. The third line is the change in conversion on the runs both cues reach and has no sign in general. The stopped comparison of Eq. (2) sees neither the first price nor D's later completions, which is why it can favour an early cue that the continued comparison rejects. Appendix J gives the proofs and the remaining identities.

5 Live Continuation under Recommended Sampling

We execute all five policies on the same 128 exposed MATH questions, batch size eight, under three fixed sampling seeds. Qwen3.5-9B (Qwen Team, 2026) runs with the same tokenizer, model implementation and decoding code for every arm, using the model card's final Best Practices sampling settings for each mode, selected before the experiment. The 2048 output cap is a deliberately tight resource regime, below the card's typical recommended allowance. Appendix Table 34 gives the verbatim prompts and forced texts, Appendix I.1 the sampling parameters, and Appendix Figure 1 summarizes the result; Appendix B lists the complete records.

Fixed policies and identity checks. D continues directly to EOS or 2048. L adds a fixed four-token final-answer instruction at 1916, retaining the prefix and leaving 128 answer tokens. T continues thinking to 2048; F uses a fixed six-token close-and-answer instruction after 1024 unless thinking has already closed. C stops D at 1024 and, without EOS, discards it and restarts T within the remaining 1024 tokens, reserving 128 answer tokens and charging the suffix. Seeds are reset by question batch and phase, arm order rotates, every intervention is part of native generation, the prefixes agree in all 1152 comparisons and all 1920 outcomes satisfy the total cap.

The late-answer control improves correct delivery. In Table 2, L raises accuracy from 80.21% to 88.02%. The registered inference unit is a question's mean over the three seeds, not 384 independent outcomes. A 20000-draw paired question-cluster bootstrap gives a 7.81-point difference with interval [4.43, 11.72]; the 99999-draw cluster sign-flip test gives Holm-adjusted $p = .00003$ over the three registered contrasts. Per-seed rescues are 14, 7 and 9, with no observed harms, on 18 unique questions. All 318 runs that finish before the late intervention exactly match D; 66 runs trigger it. Every rescue originally reached 2048 tokens without EOS or a final boxed answer; 27 of the 30 had no correct permissive extraction, and the permissive count rises from 312 to 340, so the changes exceed relabelling an unchanged answer as delivered.

Realized cost also decreases in this workload. D and L use the same 37926 prefill tokens and one native generation call each, so L has no repeated prefill. L's generated tokens, including its 264 forced tokens, are 0.981 of D's (95% batch-bootstrap interval 0.978–0.984) and its batch-policy time 0.957, interval 0.950–0.964.

D beats C at the same cap. D reaches 80.21% against C's 67.45%, so C−D is -12.76 points $[-17.71, -8.33]$ with 4 gains, 53 harms and Holm $p = .00003$, while C uses 8.69% more output and 42.69% more prefill; on the 133 escalated runs D, L and C deliver 69, 99 and 20 correct answers. The comparison concerns the restart bundle as a whole: mode switch, discarded context and smaller remaining allowance. F, which shares T's sampled prefix up to 1024 tokens, improves on T by $+46.35$

Table 3: Fixed policies on unused MATH questions with three streams each. Top: first rerun, scored with the delivery contract of Table 2; Correct over 594 runs with a unique reference, EOS and capped over 600. Bottom: preregistered second rerun on 200 further questions, 600 runs all with a unique reference, with the delivery scorer as registered primary endpoint and Qwen3.5-9B at cap 2048 unless marked; P adds a budget-aware system instruction to D. Ministral L−P is a registered descriptive contrast. G/H: gains and harms, the runs only the first or only the second policy delivers correctly. Contrasts average each question over its streams; 95% intervals use 20000 question-cluster bootstrap draws; every sign-flip $p < 10^{-4}$ except Ministral L−P at $p = .00025$.

Policy	Correct	%	EOS	Capped	Output	Contrast	Diff. [95%]	G/H
D	430	72.39	462	138	1038.5	L−D	+7.41 [4.88, 10.10]	45/1
L	474	79.80	590	10	1012.5	C−D	−8.25 [−11.95, −4.71]	17/66
T	102	17.17	102	498	1920.5	F−T	+45.79 [40.07, 51.52]	273/1
F	374	62.96	586	14	1069.9	L−C	+15.66 [11.62, 20.03]	98/5
C	381	64.14	581	19	1098.5			
D	463	77.17	492	108	972.0	L−D	+7.33 [5.00, 10.00]	45/1
L	507	84.50	590	10	951.8	L−P	+8.50 [5.67, 11.50]	65/14
P	456	76.00	491	109	968.7			
Ministral-3-8B, cap 2048, H3								
D	425	70.83	482	118	864.7	L−D	+6.83 [4.17, 9.83]	42/1
L	466	77.67	586	14	843.4	L−P	+7.33 [3.50, 11.33]	63/19
P	422	70.33	488	112	739.9			
Qwen3.5-9B, cap 4096, H4								
D	509	84.83	545	55	1226.1	L−D	+4.50 [2.50, 6.83]	27/0
L	536	89.33	592	8	1216.8			

points [39.58, 53.39] with 179 gains and one harm. Appendix B gives costs, endpoint sensitivity and seed-specific counts.

6 Replication on Unused Questions

The development matrix reuses exposed questions, so we fixed all five policies, the prompts, the seeds and the 2048-token cap, and executed them once on 200 MATH test questions that no earlier analysis had used. The questions are the first 200 of a pseudorandom ordering of the 3101 project-unexposed MATH test items, fixed before generation. Each question again receives three sampling streams, 3000 outputs in all. Two questions whose official solutions contain more than one distinct boxed value have no single reference and are excluded from correctness, leaving 594 scored runs; delivery and cost counts keep all 600 runs per arm. The top block of Table 3 scores this first rerun with the delivery contract of Table 2, and Appendix Table 35 lists the numeric-identity analysis named in its plan. The second rerun was then registered with the delivery scorer as its primary endpoint.

The delivery gain replicates. L improves correct delivery by 7.41 points, close to the 7.81 on the development set; the left panel of Appendix Figure 2 compares the two. The gain is concentrated where the mechanism predicts: all 45 rescued runs come from direct streams that reached the cap without EOS, and 32 questions improve while one worsens. L's generated tokens are 0.975 of D's and its batch-policy time 0.958.

A preregistered second rerun, and a budget-aware prompt. Before generating it, we registered MATH ranks 201–400 of the same ordering, the unchanged policy with the same cue text, position 1916, cap, sampler and seeds, the delivery scorer as primary endpoint, and a direct baseline P that appends to D's system prompt an instruction to stay within the budget and reserve room for one boxed answer; P gets the same generation cap and its 37 extra prompt tokens are not charged. In the bottom block of Table 3, L again beats D by 7.33 points [5.00, 10.00] with 45 rescues and one harm, the registered hypothesis H1, and beats P by 8.50 points [5.67, 11.50], H2. The same registration fixed two further confirmatory hypotheses on these questions: H3, L−D for Ministral-3-8B-Instruct-2512 (Mistral AI, 2025), Ministral-3-8B below, at 2048 tokens, and H4, L−D for Qwen3.5-9B at 4096 tokens, with the same cue text 132 tokens before the cap. L gains +6.83 points [4.17, 9.83]

on Ministral with 42 rescues and one harm, and $+4.50$ [2.50, 6.83] at 4096 with 27 rescues and no harm. On Ministral L also exceeds P by 7.33 points [3.50, 11.33] in a registered descriptive contrast. All four hypotheses hold after Holm correction, each with adjusted $p = 2.0 \times 10^{-4}$ in Appendix Table 35. In all three experiments the runs that end before the cue are token-identical in D and L, and L generates 2.1%, 2.5% and 0.8% fewer tokens; Appendix C gives the details.

Scorer sensitivity and delivery. The numeric-tolerant scorer also deletes unrecognized LaTeX commands, so it can accept inequivalent forms such as a brace-free fraction command against the integer 12. A comparator that never drops a command confirms every positive verdict of the first two studies, 1417 of 1426 in the preregistered Qwen rerun, 5048 of 5090 in the other comparisons of the reruns and the follow-up and 12205 of 12295 in the third study; all but three of the others are equivalent forms outside its grammar, and crediting the equivalent answers the scorer rejects leaves every L−D at or above its reported value, as Appendix Tables 10, 22 and 28 show. In the prospective fourth study it confirms 4525 of 4539 positive verdicts, and counting only confirmed deliveries leaves the 128-minus-512 reserve differences at $+14.67$ [10.83, 18.67] and $+10.17$ [6.50, 14.17] points. What the gain measures is fixed by the untreated runs: 137 of the 138 capped D outputs end without a complete final boxed field, as do 114 of 118 on Ministral and 53 of 55 at 4096 tokens, so the rescued runs gain a *delivered* correct answer. Relaxing the contract by dropping EOS, reading the last box anywhere or using a permissive read-out keeps every L−D interval above zero in Appendix Table 17.

A registered second benchmark. A registered transfer runs the unchanged policy on all 198 questions of GPQA-Diamond (Rein et al., 2024), the Diamond subset of the Graduate-Level Google-Proof Q&A benchmark, with Qwen3.5-9B, cap 2048, the cue at 1916 and three streams; the answer is one option letter in the final box, and hypothesis H5 forms its own family. Direct streams are far more exposed than on MATH: 375 of 594 runs are active at the cue, and 357 of the 359 capped D outputs contain no final box. L gains $+20.20$ points [16.50, 24.07] with 120 rescues, no harms and $p = 5.0 \times 10^{-5}$, and generates 4.1% fewer tokens, and the frozen prediction that the conversion lies in [0.20, 0.60] holds at .32, as Appendix F details. With Ministral, under H8 as its own family, 112 of 594 runs are active at the cue and L gains $+4.55$ [2.69, 6.57] with 27 rescues, no harm and $p = 5.0 \times 10^{-5}$, again with fewer tokens; Appendix H reports this study.

Registered continuations: the cue names the answer field. Two further registered studies, in Appendices E and G, re-continue every archived D stream of three second-rerun settings, and then of GPQA-Diamond, from its own token prefix through one path: prompt, prefix, forced tokens, native sampling to EOS or the cap. Their matched no-cue control N0 shares D's prefix token for token but resamples the continuation, so it reproduces D's delivery rate rather than its tokens: it delivers within 0.67 points of D at every MATH cue position we used and within 1.35 on GPQA-Diamond, and contrasts against N0 compare the forced tokens with a continuation resampled through the same path. A box-opening cue, which writes the same field label and opens the final box, beats N0 by $+7.33$ [4.83, 10.00] with Qwen at 2048 tokens under H6a, with Holm $p = 1.5 \times 10^{-4}$. Three component cues of four or five tokens at the registered trigger separate the parts of the field cue: a box opener alone, the deadline sentence *Time's up.* alone, and the label *Next step:*, which names neither an answer nor a stop. The registered cue exceeds every component in every MATH setting of Table 4: the box opener by 1.83–4.83, the deadline by 3.83–6.50 and the neutral label by 4.17–8.17 points. We therefore claim the effect for cues that name the final-answer field; with Qwen most of it is carried by the answer-field syntax.

Comparisons with adapted prior cues and budget prompts. The third study tests adapted alternatives under our accounting, and Table 4 collects the results. Our budget-charged variant of Sun et al. (2025) inserts their conclusion message 25 tokens before the cap, and the registered cue exceeds it by $+3.50$ [1.50, 5.67] with 26 gains and 5 harms under H7b, Holm $p = 1.1 \times 10^{-3}$, and by $+2.17$ and $+2.33$ points with Ministral and at 4096 tokens. Placed at our trigger instead, the message gains a further $+2.17$ [0.33, 4.17] points with Qwen at 2048 tokens and $+3.17$ at 4096, so this comparison also changes answer room: our variant leaves 11 tokens after the 14-token message, whereas the original protocol generates 25 after the message. We therefore attribute the H7b difference to the cue together with its answer allocation, not to the cue text alone. A system prompt Pn that states the numeric budget and the answer reserve in the style of Han et al. (2025), generated afresh with P's sampler and seeds, delivers 450 correct answers of 600 and generates 9% more tokens than D;

Table 4: Registered direct comparisons and controls under the delivery contract, as differences in points with 95% question-cluster intervals over 600 runs, 594 on GPQA-Diamond. Rows re-continue the archived D streams, except Pn, a fresh generation with a system prompt stating the numeric budget and the reserve. R: registered cue at the 128-token reserve, replayed on MATH and the one-pass L on GPQA-Diamond; Sun: our budget-charged variant of Sun et al. (2025)'s concluding message at $B-25$; N0: matched no-cue continuation; box opener, deadline, neutral label: component cues at the registered trigger; subscripts: answer reserve. Appendix Table 24 gives every cell. —: not in the design.

Contrast	Qwen, 2048	Ministral, 2048	Qwen, 4096	GPQA-Diamond, Qwen, 2048
Sun at $B-25$ − D	+3.50 [1.67, 5.67]	+5.00 [2.50, 7.83]	+1.83 [0.83, 3.00]	+18.52 [14.81, 22.39]
R − Sun at $B-25$	+3.50 [1.50, 5.67]	+2.17 [0.33, 4.17]	+2.33 [0.83, 4.00]	App. Table 24
L − Pn	+9.50 [6.17, 13.00]	+4.83 [0.83, 8.83]	—	—
box opener − N0	+5.83 [3.50, 8.33]	+2.50 [0.83, 4.50]	+2.17 [1.00, 3.50]	—
R − box opener	+1.83 [0.17, 3.67]	+4.83 [2.67, 7.17]	+1.83 [0.50, 3.50]	—
R − deadline	+6.50 [4.00, 9.17]	+4.67 [2.50, 7.17]	+3.83 [2.00, 6.00]	—
R − neutral label	+8.17 [5.50, 11.00]	+7.67 [4.67, 11.00]	+4.17 [2.33, 6.33]	—
R_{512} − $N0_{512}$	+3.83 [1.83, 6.00]	+4.00 [1.33, 6.83]	+3.33 [1.50, 5.33]	+17.17 [13.30, 21.21]
R_{128} − R_{512}	+3.50 [1.83, 5.33]	+2.50 [0.83, 4.33]	App. Table 29	+4.04 [1.01, 7.07]

L exceeds it by $+9.50$ [6.17, 13.00] with 70 gains and 13 harms under H7a, Holm $p = 1.0 \times 10^{-4}$. With Ministral, where Pn generates 9% fewer tokens than D, L exceeds it by $+4.83$ [0.83, 8.83] with 55 gains and 26 harms in a registered descriptive contrast.

Answer reserve: the comparison decides the choice. Against the matched no-cue control at the same position, the registered cue gains $+7.50$, $+7.67$, $+7.17$ and $+3.83$ points at reserves of 64, 128, 256 and 512 tokens with Qwen at 2048, $+6.83$, $+7.33$, $+5.67$ and $+4.00$ with Ministral, and $+4.17$, $+4.00$, $+3.17$ and $+3.33$ with Qwen at 4096, so the gap at 512 also appears against a control that shares the continuation path. Proposition 3 splits it: of the 128-token reserve's 21-run advantage, 2 runs lie among the 31 runs that end between the two positions, 27 of which D delivers correctly, and 19 among the runs both cues reach, so the later cue mainly converts more of the runs the two share. On GPQA-Diamond with Qwen, delivery again falls as the reserve grows, with 338, 333, 315 and 309 correct at 64–512 tokens; Appendix Tables 20 and 27 give the frozen forecasts. A fourth registered study, reported in Appendix H, then tested the choice prospectively in operating conditions where no reserve outcome had been generated. At a 1024-token cap the 128-token reserve beats 512 by $+14.33$ points [10.67, 18.33] with Qwen, 93 gains and 7 harms under H10a, and by $+10.33$ [6.50, 14.33] with Ministral, 67 gains and 5 harms under H10b, each with Holm $p = 1.5 \times 10^{-4}$ over the three planned members; both hold under every read-out, and N0 at 512 delivers within 0.67 points of D. The 512-token reserve generates 26–29% fewer tokens than the 128-token one. The 64-token reserve, which three of four frozen predictors including the constant-gain rule chose, delivered most in all three new conditions, $+16.00$ and $+12.00$ points over 512 at 1024 tokens. When correct delivery is the objective, the evidence therefore favours a short reserve of 64–128 tokens over 512: $R_{128}-R_{512}$ is positive in all four 2048- and 1024-token MATH settings and on GPQA-Diamond with Qwen under each of three read-outs in Appendix Tables 29 and 31, within which 64 and 128 tokens lie close together.

Where direct continuation beats restarting. Direct continuation again beats restart C by $+8.25$ points: 66 runs only D delivers correctly and 17 only C. In the 96 restarted runs whose direct stream would have finished before 1916 tokens, restart destroys 62 correct deliveries and recovers one, and none of the 239 restarted drafts held a complete boxed answer at 1024 tokens; Appendix Table 14 gives the partition.

7 Discussion and Conclusion

When a late instruction can help. By Proposition 2, L−D is the share of direct streams still active at the cue times the *conversion*, the conditional net gain, gains minus harms over active runs. Table 5 lists both for every executed setting, and Appendix D analyses them across settings. They support one scoped claim, for MATH with two model families at caps of 1024–4096 tokens and GPQA-Diamond

Table 5: Exposure and conversion in every executed setting under the delivery contract, on MATH unless marked. Cap-hit: D runs without EOS at the cap. Active: D runs without EOS at the cue, $B - 132$, the only runs L can change, with their share. Conv.: conversion, the net gain over active runs. Read-outs: range of L−D over the four answer read-outs of Appendix Table 17, or over the three registered read-outs of Appendix F on GPQA. G/H: gains and harms. Tokens: L/D generated tokens, forced tokens included. Runs per arm: 384 in development, 594 in the first rerun and on GPQA, 600 in the second rerun. The 1024-token settings are registered secondary experiments.

Model, cap	Study	Cap-hit	Active	D	L	L−D [95%]	G/H	Conv.	Read-outs	Tokens
Qwen, 2048	development	15.4%	66 (17%)	308	338	+7.81 [4.43, 11.72]	30/0	.45	7.29–8.07	.981
Qwen, 2048	first rerun	22.9%	140 (24%)	430	474	+7.41 [4.88, 10.10]	45/1	.31	6.23–7.58	.975
Qwen, 1024	second rerun	36.3%	245 (41%)	366	422	+9.33 [6.50, 12.33]	58/2	.23	8.67–9.33	.939
Qwen, 2048	second rerun, H1	18.0%	121 (20%)	463	507	+7.33 [5.00, 10.00]	45/1	.36	6.67–7.33	.979
Qwen, 4096	second rerun, H4	9.2%	55 (9%)	509	536	+4.50 [2.50, 6.83]	27/0	.49	4.50–4.83	.992
Ministral, 1024	second rerun	35.7%	232 (39%)	341	402	+10.17 [6.50, 14.17]	62/1	.26	10.17–11.83	.941
Ministral, 2048	second rerun, H3	19.7%	129 (22%)	425	466	+6.83 [4.17, 9.83]	42/1	.32	6.00–7.00	.975
Qwen, 2048	GPQA-Diamond, H5	60.4%	375 (63%)	213	333	+20.20 [16.50, 24.07]	120/0	.32	20.20–24.24	.959
Ministral, 2048	GPQA-Diamond, H8	18.5%	112 (19%)	241	268	+4.55 [2.69, 6.57]	27/0	.24	4.55–4.88	.981

with both at 2048, this cue and a 128-token reserve: L−D is positive in all nine settings, from $+4.50$ to $+20.20$ points, and under every registered answer read-out, every change lies in an active run, and 23–49% of the active runs are converted net of harms. Exposure, which D alone reveals, varies more than fourfold; D's cap-hit rate and L−D have rank correlation .80 across the five second-rerun settings in a registered descriptive analysis over settings that share questions. Conversion rises as the cap loosens, for Qwen .23, .36 and .49 at 1024, 2048 and 4096 tokens and for Ministral .26 and .32, so exposure bounds the gain without fixing it.

Where the cue rescues. Among the 915 active MATH runs that D does not deliver correctly, L rescues 189 of the 262 in which a checker match finds the correct value, always alongside incorrect candidates, but 120 of the other 653: 72% versus 18%, a difference of $+0.54$ [0.43, 0.63]. The audit records what was written, not what the model had adopted; Appendix Table 18 gives the counts.

Implications for evaluating answer-stage interventions. The comparison decides the conclusion. Against the stream stopped at the cue, Eq. (2), a 512-token reserve looks best in the three follow-up MATH settings; against the stream continued to the same cap, Eq. (3), each shorter reserve beats it in both 2048-token and both 1024-token settings, as Appendices G and H report. Compare an intervention with its own stream continued to the same total cap; score delivery alongside extraction, since they disagree where interventions act, and 27 of 30 development rescues had no correct permissive extraction; and charge every forced, discarded and regenerated token and repeated prefill.

Scope. The evidence concerns MATH with Qwen3.5-9B at 1024, 2048 and 4096 total tokens and Ministral-3-8B at 1024 and 2048, and GPQA-Diamond with both models at 2048; the four-token cue with a 128-token answer reserve, and in the follow-ups reserves of 64–512 tokens, a box-opening variant, three component cues and budget-charged variants of Sun et al. (2025)'s message without its prompt notice; and two system-prompt baselines. Cue effects are conditional on the fixed output contract, whose system prompt already names the final-answer field that the cue repeats. L needs only an output limit and an insertion point, or else two calls and one extra prefill. Correct delivery is acceptance under the fixed, numeric-tolerant scoring contract, not a complete exact semantic audit. That contract counts a rejected answer and a missing one alike; where a wrong answer costs more than none, delivered wrong answers should be weighed as well, as the GPQA-Diamond counts of Appendix F allow.

Conclusion. A late cue that names the final-answer field improves correct delivery at the same total cap with slightly less work on MATH and GPQA-Diamond, in two model families and at caps of 1024–4096 tokens; the gain is confined to runs active at the cue, and the best answer reserve depends on whether the comparison stops the stream at the cue or continues it to the same cap; against the continued stream a short reserve beat a 512-token one in a prospective test. Continuing the direct stream also beats restarting thinking.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. Formulating questions for surveys or interviews and transcribing recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study evaluates openly released model weights on the public MATH and GPQA-Diamond benchmarks. It involves no human subjects, no personal or sensitive data and no deployment to users. The supplementary archive releases model outputs on these public questions; it includes the MATH questions and reference solutions as distributed, whereas GPQA-Diamond items are referenced by identifier and rebuilt from the public release, so that the benchmark's questions are not redistributed in plain text. The interventions studied change only when a model is asked to state its answer, so we see no specific dual-use or safety risk beyond those of the underlying models; answer forcing can make an unfinished solution look final, which is why we score delivery separately from extraction and report delivered wrong answers. No fairness-relevant populations or individual-level data are involved.

REPRODUCIBILITY STATEMENT

Section 3 and Table 1 define the policies, endpoints and total-cap accounting; Section 5, Appendix I and Table 34 give the models, prompts, chat templates, sampling parameters, seeds and forced texts; Appendices C, D, E, F, G and H describe the reruns, the registered follow-ups and transfers and the prospective reserve test with their registrations and analyses; Section 4 and Appendix J state the assumptions and give complete proofs; Appendix K lists every registered test and frozen prediction. The anonymous supplementary archive contains the generation and scoring code, the registration documents, the questions and references, the per-run records and the scripts that recompute the reported tables, indexed by claim in Appendix A; its reproduction guide describes three levels of reproduction: recomputing every count and paired contrast from the per-run outcome table, rescoring the archived outputs, and regenerating outputs with the released code, prompts and seeds.

REFERENCES

Francesca Carlon, Vincent Ginis, and Andres Algaba. Shorter reasoning, earlier answers? An evaluation of reasoning interfaces. *arXiv preprint arXiv:2608.03401*, 2026. URL https://arxiv.org/abs/2608.03401.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bx.html.

Ayan Datta, Zhixue Zhao, Bhuvanesh Verma, Radhika Mamidi, Mounika Marreddy, and Alexander Mehler. Large language models decide early and explain later. *arXiv preprint arXiv:2604.22266*, 2026. URL https://arxiv.org/abs/2604.22266.

Zhe Dong, Fang Qin, and Manish Shah. When does learning to stop help? A cost-aware study of early exits in reasoning models. *arXiv preprint arXiv:2606.30852*, 2026. URL https://arxiv.org/abs/2606.30852.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling LLM reasoning programs with Certaindex. In *Advances in Neural Information Processing Systems*, vol-

ume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/d037fd021c9aace128b8ce25001cdb6c-Abstract-Conference.html.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. Inverse scaling in test-time compute. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=NXgyHW1c7M.

Rodrigo Guedes de Souza and Alison R. Panisson. Who thinks best depends on how long you let them: Budget-dependent rankings in LLM evaluation. *arXiv preprint arXiv:2608.12150*, 2026. URL https://arxiv.org/abs/2608.12150.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1274. URL https://aclanthology.org/2025.findings-acl.1274/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Parsa Hosseini, Sumit Nawathe, Mahdi Salmani, Meisam Razaviyayn, and Soheil Feizi. Early stopping for large reasoning models via confidence dynamics. *arXiv preprint arXiv:2604.04930*, 2026. URL https://arxiv.org/abs/2604.04930.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Jungseob Lee, Seongtae Hong, Seungjun Lee, Jaehyung Seo, Junyoung Son, Sugyeong Eo, Chanjun Park, Hyeongju Park, Hyeonseok Moon, and Heuiseok Lim. DART: Draft-agreement routing for training-free adaptive thinking budgets in hybrid reasoning models. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. URL https://arxiv.org/abs/2606.23181. To appear.

Xin Liu and Lu Wang. Answer convergence as a signal for early stopping in reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17896–17907, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.904. URL https://aclanthology.org/2025.emnlp-main.904/.

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. URL https://arxiv.org/abs/2504.09858.

Mistral AI. Ministral-3-8B-Instruct-2512 model card. Hugging Face model repository, 2025. URL https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Yansong Ning, Mianpeng Liu, Jingwen Ye, Weidong Zhang, and Hao Liu. HRBench: Benchmarking and understanding thinking-mode switch strategies in hybrid-reasoning LLMs. *arXiv preprint arXiv:2605.28398*, 2026. URL https://arxiv.org/abs/2605.28398.

Renuka Oladri, Niveda Jawahar, and Abdirisak Mohamed. Token budget saturation and mechanistic early detection of reasoning non-convergence in chain-of-thought models. *arXiv preprint arXiv:2607.21433*, 2026. URL https://arxiv.org/abs/2607.21433.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Optimizing anytime reasoning via budget relative policy optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/21efcb6a9fe5c90d660d3f560bb6094d-Abstract-Conference.html.

Qwen Team. Qwen3.5-9B model card. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Ian Su, Gaurav Purushothaman, Jey Narayan, Ruhika Goel, Kevin Zhu, Sunishchal Dev, Yash More, and Maheep Chaudhary. Broken chains: The cost of incomplete reasoning in LLMs. *arXiv preprint arXiv:2602.14444*, 2026. URL https://arxiv.org/abs/2602.14444.

Yi Sun, Han Wang, Jiaqiang Li, Jiacheng Liu, Xiangyu Li, Hao Wen, Yizhen Yuan, Huiwen Zheng, Yan Liang, Yuanchun Li, and Yunxin Liu. An empirical study of LLM reasoning ability under strict output length constraint. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7652–7671. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.389. URL https://aclanthology.org/2025.emnlp-main.389/.

Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. Demystifying hybrid thinking: Can LLMs truly switch between think and no-think? *arXiv preprint arXiv:2510.12680*, 2025. URL https://arxiv.org/abs/2510.12680.

Tsung-Han Wu, Mihran Miroyan, David M. Chan, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. Are large reasoning models interruptible? In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=mlQWk1DtkN.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. Scalable chain of thoughts via elastic reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=E0Qfhma53J.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://openreview.net/forum?id=NpU7ZXafRi.

Van Yang, Shouren Wang, Debargha Ganguly, Xinpeng Li, Chaoda Song, Vikash Singh, Vipin Chaudhary, and Xiaotian Han. Mid-Think: Training-free intermediate-budget reasoning via token-level triggers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6024–6038. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.299. URL https://aclanthology.org/2026.findings-acl.299/.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. AdaptThink: Reasoning models can learn when to think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3716–3730. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.184. URL https://aclanthology.org/2025.emnlp-main.184/.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in LLM test-time compute scaling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23967–23977. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1199. URL https://aclanthology.org/2026.findings-acl.1199/.

A CLAIM INDEX

Table 6 maps each main-text claim to the study that supports it, its endpoint and its evidentiary role. The reproduction guide of the supplementary archive maps each claim to its input records and recomputation script. We call the second rerun preregistered because its questions, policies, endpoint and analysis were fixed in writing before any of its generations; a registered study or test is one whose hypotheses, predictions and analysis were fixed in writing before its outputs were generated or, where stated, before they were read.

Table 6: Claims, supporting studies, endpoints and evidentiary roles.

Claim	Evidence	Endpoint	Role
L beats D at the same 2048 cap, +7.81	Table 2, 128 questions	delivery	development
The gain replicates, +7.41	Table 3, 200 unused questions	delivery	fixed-policy replication
The gain replicates again, +7.33, and exceeds a budget-aware prompt, +8.50	Table 3, App. C	delivery, registered primary	preregistered replication
The gain holds across scorers	Tables 10, 22 and 28, App. B.3 and the fourth-study audit of App. H	four scorers; command-preserving audit	development, reruns, follow-up, third and fourth studies
The identified effect is delivery: 137 of 138 capped D outputs lack a complete final box	Section 6	boundary fields	replication
The gain holds for Ministral-3-8B and at 4096 tokens, H3 and H4	Table 3, App. C	delivery, registered primary	preregistered replication
Restart damage falls on runs D would finish	Table 14	delivery	replication
Exposure bounds the gain by level	App. Table 15, Prop. 2	delivery	both reruns
The gain transfers to GPQA-Diamond, H5, +20.20	Section 6, App. F	delivery, letter	registered transfer
The gain transfers to GPQA-Diamond with Ministral-3-8B, H8, +4.55	Section 6, App. H	delivery, letter	fourth registered study
L−D is positive in all nine settings and under every registered read-out; conversion 23–49%	Table 5, App. D	delivery; read-outs	all studies; secondaries registered
Rescues concentrate where a checker match finds the correct value before the cue	Section 7, App. Table 18	delivery; pre-cue text	all studies, descriptive
Box-opening cue confirmed under H6a; answer-field cues exceed prose stop cues; frozen reserve predictions hold	App. Tables 19, 20 and 21	delivery	registered follow-up
The cue exceeds the prior concluding message under H7b and the numeric-budget prompt under H7a, and exceeds each component cue	Table 4, App. Tables 24 and 25	delivery; box without EOS	third registered study
The reserve ranking reverses between the stopped and the continued comparison; N0 at every reserve	App. Table 26, Prop. 3	delivery	third registered study, descriptive
Held-out reserve prediction on GPQA-Diamond	App. Table 27	delivery, letter	third registered study, frozen predictions
A 128-token reserve beats 512 in two new 1024-token conditions, H10a and H10b; the 64-token reserve chosen by three of the four frozen predictors delivers most	App. Tables 31, 32 and 33	delivery; three read-outs	fourth registered study, prospective
L uses fewer tokens in all studies and less batch time in the first two	Sections 5–6	cost	all studies

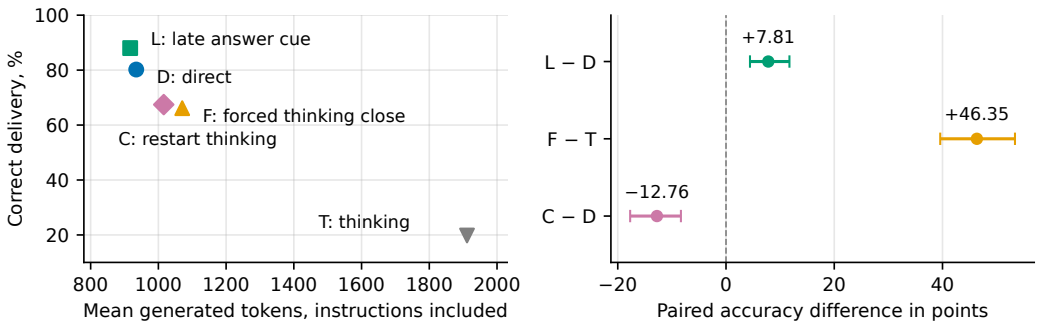

Figure 1: **At a matched output allowance, the late answer cue improves correct delivery and outperforms restart.** Left: all five arms' delivered accuracy and realized mean output; lower output and higher accuracy are preferable. Right: registered paired accuracy differences with 95% question-cluster intervals, averaging each of 128 questions over its three fixed seeds. L and C compare with D; F compares with T. All outcomes, including caps, remain in the denominator.

B Complete Live Sampling Evidence

This appendix gives the complete development matrix of Section 5; Figure 1 summarizes it, and Appendix I gives the design, sampling parameters and forced texts.

B.1 Endpoints, uncertainty and complete matrix

The primary endpoint requires natural EOS and a complete last balanced boxed answer in the answer region, with no later unclosed box or substantive trailing text. Whitespace, math or Markdown delimiters and a terminal full stop are allowed. Thinking without its generated closing token cannot deliver. The box is evaluated by the unchanged MATH scorer, which includes numeric tolerance; acceptance does not imply exact equality. The permissive view uses the earlier last-number parser and does not require EOS or a final field. These endpoints answer different questions.

Question-level inference averages the three binary outcomes before resampling. The 20000 paired bootstrap draws use a fixed seed, and the reported endpoints are the 500th and 19500th of the sorted draws. The subsequent 99999 Monte Carlo sign flips of the 128 mean differences give two-sided descriptive p values, assuming sign symmetry under the null, with add-one correction. Holm adjusts the three prespecified contrasts. Intervals condition on the exposed question pool and these three streams. Per-seed exact McNemar values in Table 8 are secondary and not a second selected family.

B.2 Retained-prefix transitions and failure analysis

All 318 D runs ending before the late intervention are token-for-token identical under L. The remaining 66 runs cover 30 distinct questions: D delivers seven correct answers and L 37; EOS increases from seven to 61 and permissive correctness from ten to 38. All 30 primary rescues reach D's 2048 cap without EOS or a final boxed answer; under L they have a correct final box and EOS. Three already had a correct permissive D extraction, while 27 did not. The full permissive contrast has 28 gains and zero harms, rather than 30, because its contract differs.

The 30 primary gains involve 18 questions: three improve under all three seeds, six under two seeds, nine under one; the other 110 question means are unchanged. A nonblind reading of all rescued tails finds many correct values already in the direct trace before additional checking. Thus a wrong last-number extraction is not proof of failed reasoning; the intervention acts on the reminder, answer selection, format and continuation together, and its measured outcome is delivery.

C restarts on 133 question-seed runs over 53 questions. On that same subset, correct deliveries are D 69, L 99 and C 20, and D alone delivers correctly in 53 runs, C alone in four. Its 251 nonescalations exactly equal D. The subset comparison is descriptive; the full-denominator cluster comparison is the primary result. F improves on T and leaves nine runs without EOS, nine without a box and seven

Table 7: Every seed and arm, each denominator 128. Seeds 20, 21 and 22 label the three fixed sampling streams. Output counts all generated and injected tokens; prefill counts actual, unpadded prompt tokens; active counts rows receiving a suffix. Wall is summed batch-policy seconds.

Seed	Arm	Correct	EOS	Output	Prefill	Active	Wall
20	D	100	106	121079	12642	0	1560.20
20	L	114	125	118738	12642	25	1495.95
20	T	27	27	243956	12386	0	1654.12
20	F	85	126	135825	12386	115	996.89
20	C	90	125	127921	17930	43	1550.21
21	D	103	109	118257	12642	0	1542.81
21	L	110	126	115985	12642	22	1477.67
21	T	26	26	243762	12386	0	1657.20
21	F	83	126	136069	12386	116	1054.80
21	C	84	125	130653	18088	45	1507.38
22	D	105	110	119563	12642	0	1511.91
22	L	114	128	117300	12642	19	1440.85
22	T	23	23	246256	12386	0	1661.23
22	F	86	123	139155	12386	117	1166.06
22	C	85	123	131504	18097	45	1529.56

Table 8: All fixed primary contrasts by seed. Gains and harms are paired question counts, not independent observations across seeds. Exact two-sided McNemar values are unadjusted secondary descriptions.

Contrast	Seed	Gains	Harms	p
C−D	20	2	12	0.0129
	21	1	20	2.1×10^{-5}
	22	1	21	1.1×10^{-5}
F−T	20	58	0	6.94×10^{-18}
	21	58	1	2.08×10^{-16}
	22	63	0	2.17×10^{-19}
L−D	20	14	0	0.000122
	21	7	0	0.0156
	22	9	0	0.00391

with substantive trailing text. T has 308 runs without EOS, including 285 without a thinking close. Failure types are retained rather than filtered.

Table 9 retains all ten pair counts. The unregistered L−C contrast is 20.57 points with cluster interval [14.58, 26.82], 81 gains and two harms. L uses less actual output and prefill and 3.76% less wall than C, but 27.82% more padded visits. This contrast is exploratory and outside the three-contrast Holm family.

Table 9: All pairwise outcome counts across 384 paired runs. These descriptive counts are not tests treating 384 runs as independent questions.

Contrast	Gains	Harms	Difference, points
D−L	0	30	−7.8125
D−T	232	0	+60.4167
D−F	62	8	+14.0625
D−C	53	4	+12.7604
L−T	262	0	+68.2292
L−F	87	3	+21.8750
L−C	81	2	+20.5729
T−F	1	179	−46.3542
T−C	0	183	−47.6562
F−C	22	27	−1.3021

Development references. The development scorer uses gold values fixed before the study; they equal the unique boxed value of the official solution for 127 of the 128 questions. For one precalculus question that asks for all possible values, the official solution boxes -1 and 2 and the stored gold is 2. D and L deliver the complete pair in all three streams, so scoring that question against the official answer gives D, L, T, F and C 311, 341, 76, 254 and 259 correct runs and leaves L$-$D at $+7.81$ with 30 gains and no harm. The main text keeps the frozen references.

B.3 Command-preserving audit of the MATH scorer

Table 10: L$-$D under four scorers, in percentage points. All four apply the same delivery contract and differ only in how an answer is judged equivalent to the reference; Table 5 gives robustness to the contract itself. Development: 384 runs. Unused questions of the first rerun: 594 runs for the first two rows; the numeric-identity bounds and the semantic labels use all 600 runs, with the two questions lacking a unique reference left unresolved. Bounds are partial-identification ranges, not confidence intervals.

Scorer	Development, 128 questions	Unused, 200 questions
Paper delivery scorer	338 vs 308: $+7.81$	474 vs 430: $+7.41$
Command-preserving equivalence check	342 vs 312: $+7.81$	482 vs 438: $+7.41$
Numeric-identity comparator, identity-consistent bounds	[$+6.77$, $+8.85$]	[$+5.33$, $+13.00$]
Model-assisted semantic adjudication	$+7.81$ on 42 focal outputs	480 vs 435: [$+4.17$, $+11.83$]

The MATH scorer used for every delivery verdict normalizes both strings, rewrites fraction and root commands into brace-counted forms, and then deletes any remaining LaTeX command and all braces before comparing strings, numbers and, as a fallback, symbolic expressions. Its numeric branches accept $|p-g| \le 10^{-4}\max(1,|g|)$; its matrix branch permits an absolute difference of 10^{-6} per entry. These rules define the reported operational endpoint rather than exact semantic equality. The deletion step also makes it accept inequivalent pairs whose remaining characters coincide: a brace-free fraction of 1 over 2 against 12, a binomial coefficient of 5 and 2 against 52, and the sine of 1 against 1. We therefore audited every positive verdict of both matrices with a separate comparator that never drops a command. It accepts a pair only if the strings agree after formatting-only normalization of spaces, dollar signs, delimiter-sizing and spacing commands, display fractions, text wrappers, degree marks, a leading variable assignment and a trailing period, or if both sides translate into symbolic expressions through an explicit translator for fractions, roots, binomials, trigonometric and logarithmic functions, π and powers, which leaves any other command unresolved, and the difference simplifies to zero. It uses no numeric tolerance and does not drop percent signs. Forms it cannot translate stay unresolved and are never credited.

All 1235 positive verdicts of the development matrix, 1210 by normalization and 25 symbolically, and all 1761 of the first rerun over the 594 scored runs per arm, 1686 by normalization and 75 symbolically, are confirmed; none depends on the deletion step. In the reverse direction, 14 development and 28 confirmation outputs are delivered and equivalent under the comparator but rejected by the scorer, mostly brace-free fractions or degree marks written against a reference in another notation. Crediting them gives D, L, T, F and C counts of 312, 342, 76, 255 and 264 in development and 438, 482, 102, 379 and 388 in the first rerun. The registered contrasts become L$-$D $+7.81$ [4.43, 11.72] with 30 gains and no harm, and $+7.41$ [4.88, 10.27] with 45 gains and one harm; C$-$D -12.50 and -8.42; F$-$T $+46.61$ and $+46.63$, each from 20000 question-cluster draws. The supplementary archive contains the audit script and regression tests for the three counterexamples and for two behaviours the audit must not have, accepting 1 against 1.0000000005 and 50 percent against 50. Appendix C applies the same audit to the second rerun.

B.4 Endpoint sensitivity across all policies

Table 11 gives every arm's count under each endpoint, with denominator 384.

B.5 Resource accounting and measurement boundary

Table 12 exposes both actual token counts and padded scheduled visits, and Table 13 gives the cost ratios of the registered comparisons. Padded decode is the batch width times the number of executed

Table 11: Full endpoint sensitivity, denominator 384 for every count. Numeric: permissive correctness; EOS+numeric: both; field: complete final field and acceptance by the fixed scorer without requiring EOS. Primary additionally requires EOS.

Arm	Primary	Numeric	EOS+numeric	Field	No EOS
D	308	312	309	308	59
L	338	340	339	338	5
T	76	111	76	76	308
F	254	260	260	254	9
C	259	261	261	259	11

decode steps, including finished rows; it is a work proxy, not a count of floating-point operations. A restarted stage may change padding efficiency, so output, padded visits and wall need not move in the same direction. Generation time is synchronized. Policy wall also includes the intervention and per-arm bookkeeping; shared prompt tokenization, model loading, warmup and endpoint decoding and grading are outside that clock. Summed policy time is 21806.84 seconds. These are serial sums over batches, not single-request latency or simultaneous-throughput estimates.

Table 12: Complete resource accounting for the experimental matrix. Forced tokens are included in output, never added free. Pad-P and Pad-D: padded prefill and decode visits; actual: output plus actual prefill. Gen: synchronized generation seconds; s/correct: policy seconds per correct delivery, an aggregate throughput ratio.

Arm	Forced	Pad-P	Pad-D	Actual	Gen, s	s/correct
D	0	67272	744264	396825	4614.79	14.98
L	264	67272	712224	389949	4414.38	13.06
T	0	66504	786432	771132	4972.39	65.43
F	2088	66504	508816	448207	3217.64	12.67
C	798	92306	517519	444193	4587.01	17.71

Table 13: Ratios of fully executed costs for the three registered comparisons; less than one means less of that resource. No ratio substitutes for a different cost axis.

Ratio	Output	Prefill	Output+prefill	Padded visits	Wall
C/D	1.0869	1.4269	1.1194	0.7514	0.9940
F/T	0.5600	1.0000	0.5812	0.6745	0.6471
L/D	0.9808	1.0000	0.9827	0.9605	0.9566

L decreases output and policy wall in each of the three streams separately, as Table 7 shows, and the paired question outcomes improve in all three. D and L have exactly the same 37926 actual prefill tokens and retain the same cache. L's 264 forced tokens replace stochastic outputs within the same allowance, and its total output is 6876 tokens lower. C uses 42.69% more actual prefill and 8.69% more output than D, while its padded visits are 24.86% lower and measured wall is 0.60% lower. Therefore the resource cost of C is specifically extra actual output and prefill with lower delivered accuracy.

C Preregistered Second Rerun and Registered Analyses

Registration. The second rerun was registered in writing before any of its generations. Questions: ranks 201–400 of the fixed seeded pseudorandom ordering of the 3101 project-unexposed eligible MATH test items whose ranks 1–200 formed the first rerun, which the ordering reproduces exactly; every official solution has one distinct boxed value; level mix 19/32/48/43/58 questions at Levels 1–5. Policy: unchanged from the development study, with the non-thinking chat template and math system prompt, the four-token suffix of Table 34 forced through native generation at position $B - 128 - 4 = 1916$, total cap $B = 2048$ including forced tokens, the sampler of Appendix I.1,

the same three sampling seeds, batch size eight and rotating arm order. P appends to the system prompt the budget instruction of Table 34. The same registration fixed H3, L−D for Ministral-3-8B-Instruct-2512 with arms D, L and P at $B = 2048$, and H4, L−D for Qwen3.5-9B at $B = 4096$ with the cue at $4096 - 132 = 3964$, on the same questions, prompts and seeds. Ministral runs from the provider's released checkpoint, on text only, at the recommended temperature 0.05 without top-k or top-p truncation or penalties. Primary endpoint: the delivery contract of Section 3 with the paper scorer against the official boxed reference. Inference: question means over the three streams, 20000 question-cluster bootstrap draws and 20000 sign flips with a fixed seed, and Holm correction over the four registered hypotheses of Table 35. The registered analysis script was fixed before any outcome was viewed. It was run for H1 and H2 after all 75 units, 25 batches of eight questions in each of three sampling streams, of the Qwen3.5-9B experiment at 2048 tokens had finished, and once over H1–H4 after the 75 units of each of the Ministral and 4096-token experiments had finished; H1 and H2 are identical in both runs, and no unit was rerun.

Second-rerun details. All 75 units satisfied the D/L prefix identity and the total cap. Outside the active set the 479 L outputs are token-identical to D's; inside it D delivers 11 correct answers and L 55, with 45 rescues and one harm, a conversion of 36%. By question, 33 improve and none worsens under L−D; under L−P 41 improve and 3 worsen. Of the 108 capped D outputs, 107 contain no complete box and one has substantive text after its last box. Generated tokens: D 583210, L 571079 including 484 forced, P 581249; L/D is 0.979, a 2.08% reduction with question-batch bootstrap interval 0.975–0.984. P's prompt is 37 tokens longer for every question. Token counts are the registered cost measure of this rerun. The command-preserving check, without tolerance or percent removal, confirms 1417 of 1426 positive verdicts, 1382 by format and 35 symbolically; the nine others are a word answer written in a capitalized text wrapper and a column vector written with fraction commands against slashes, outside its grammar and equivalent by inspection. Crediting its 34 equivalent but rejected deliveries gives L−D +7.67 [5.33, 10.33] with 47 gains and one harm.

Ministral-3-8B and the 4096-token cap. All 75 units of each experiment satisfied the D/L prefix identity and the total cap, and every L suffix starts at the cue position, 1916 or 3964. Outside the active set the 471 and 545 L outputs are token-identical to D's. Inside it D delivers 9 correct answers and L 50 on Ministral, with 42 rescues and one harm as 23 questions improve and none worsens, and 0 and 27 at 4096, with 27 rescues and no harm as 19 questions improve and none worsens. Of the 118 capped Ministral D outputs, 114 contain no complete box, three have substantive text after their last box and one reaches the cap after a complete box without EOS; of the 55 capped D outputs at 4096, 53 contain no complete box and two have trailing text. L's generated tokens are 0.975 of D's on Ministral, with question-batch bootstrap interval 0.971–0.980, and 0.992 at 4096, with interval 0.991–0.995. By MATH level, the 5, 7 and 43 active runs at 4096 lie at Levels 3, 4 and 5, where L gains 5, 3 and 19 correct runs.

Table 14: Restart root cause on the 594 scored runs of the unused questions. Groups are defined by the direct stream D alone; within each group D, L and C share the same prefix and seed.

Direct stream D	Runs	D	L	C	Output D / C
ends by 1024, no restart	358	341	341	341	193864 / 193864
restarted; D ends by 1916	96	85	85	24	134621 / 185901
restarted; D active at 1916	140	4	48	16	286513 / 271511

Exposure by level. Table 15 gives the level breakdown of both reruns on unused questions. In the first rerun the active share ranges from 1 of 93 runs at Level 2 to 85 of 168 at Level 5, and L's gain from +1 to +22 correct runs, while D's advantage over restart is largest on the hardest level; in the second rerun Level 5 holds 85 of the 121 active runs and +32 of the +44 net gain. Figure 2 summarizes the first rerun's contrasts and the restart partition of Table 14.

Table 15: Exposure bounds the late-instruction gain across MATH levels in both reruns on unused questions, under the delivery endpoint, with 594 scored runs in the first rerun and 600 in the second. Active: direct streams without EOS at 1916 tokens, the only runs L can change.

	First rerun						Preregistered second rerun				
Level	Runs	Active	D	L	L−D	C	Runs	Active	D	L	L−D
1	60	9	51	58	+7	55	57	1	54	54	0
2	93	1	84	85	+1	81	96	6	84	87	+3
3	135	15	117	122	+5	110	144	13	133	137	+4
4	138	30	100	109	+9	85	129	16	103	108	+5
5	168	85	78	100	+22	50	174	85	89	121	+32
All	594	140	430	474	+44	381	600	121	463	507	+44

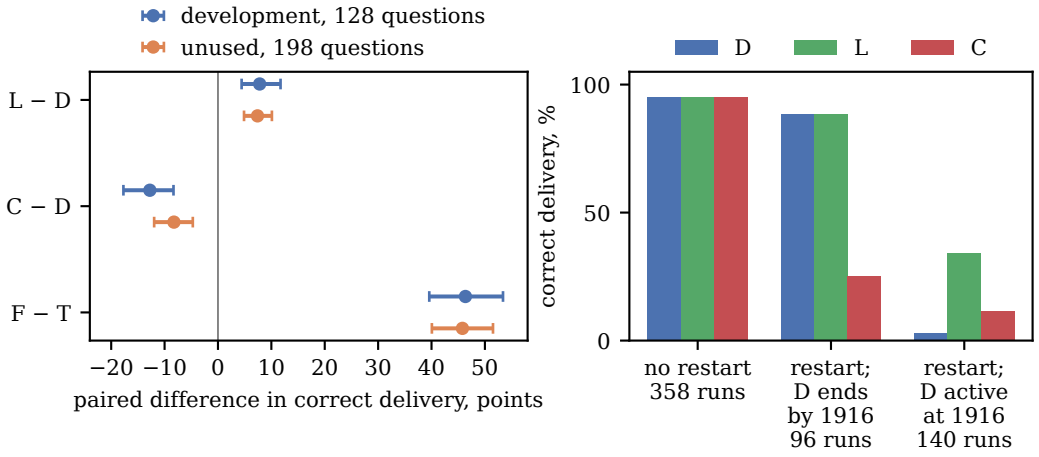

Figure 2: **The fixed policies replicate on unused questions, and D's advantage over restart lies where the direct stream would have finished.** Left: paired delivery-endpoint contrasts with 95% question-cluster intervals on the development questions of Table 2 and on the 198 uniquely referenced unused questions of the first rerun in Table 3. Right: correct delivery on those unused questions, grouped by the direct stream's own trajectory as in Table 14.

D Cross-Setting Analyses of the Executed Records

The analyses of this appendix read only records already generated for the studies above; none required a new generation. Their scripts and outputs are in the supplementary archive.

Registered secondary experiments. The registration of the second rerun also fixed secondary experiments, analysed with the same frozen analysis script once all 75 units of each of its six experiments had finished: L−D for Qwen3.5-9B and for Ministral-3-8B at $B = 1024$, with the cue at $1024 - 132 = 892$, on the same 200 questions, prompts and seeds, and the budget-aware prompt P generated on the first rerun's 200 questions and compared by question and stream with that rerun's archived L. Table 16 lists these contrasts. All 150 D/L units of the two 1024-token experiments satisfy the prefix identity and the total cap; outside the active set the 355 and 368 L outputs are token-identical to D's; of the 218 and 214 capped D outputs, 216 and 214 contain no complete box. L ends with EOS in 590 and 556 runs at 1024 tokens, against 382 and 386 for D, with L/D generated tokens of 0.939 and 0.941; on the first rerun's questions P ends with EOS in 469 of 600 runs and 131 are capped, and L generates 0.977 of P's tokens. P does not share D's sampled prefix, so L−P is not localized to active runs. The registration also stated a descriptive mechanism prediction: across settings, L−D rises with D's cap-hit rate and is near zero where that rate is below 2%. Across the five second-rerun settings of Table 5 the Spearman correlation between cap-hit rate and L−D is .80;

all five settings have cap-hit rates of 2% or more, so the correlation covers the rising part of the prediction.

Table 16: Registered secondary contrasts of the second rerun under the delivery endpoint, with question means over three streams and 20000 question-cluster bootstrap draws and sign flips, as registered.

Experiment	Contrast	Correct	Diff. [95%]	G/H	Sign-flip p
Qwen3.5-9B, B=1024	L−D	422 vs 366 / 600	+9.33 [6.50, 12.33]	58/2	5.0×10^{-5}
Ministral-3-8B, B=1024	L−D	402 vs 341 / 600	+10.17 [6.50, 14.17]	62/1	5.0×10^{-5}
Qwen3.5-9B, B=2048, first rerun	L−P	474 vs 438 / 594	+6.06 [3.03, 9.43]	50/14	3.5×10^{-4}

Answer read-outs. Table 17 scores each L and D output under four read-outs of the same text: E1, the primary delivery contract of Section 3; E2, the same final field without requiring EOS; E3, the last complete box anywhere in the answer region; and E4, the permissive read-out that takes the last box, else a "Final answer:" tail, else the last number, with EOS ignored. These differ from the four scorers of Table 10, which all apply E1 and differ only in how an answer is judged equivalent to the reference. L−D is positive with an interval excluding zero under every read-out in every setting. Among the 309 E1 rescues, D's text before the cue already contained a complete correct box in 3, two at 4096 tokens and one on Ministral at 2048, and a correct permissive read-out in 21.

Table 17: L−D in points under four answer read-outs of the same outputs, with 95% question-cluster bootstrap intervals from 20000 draws. E1 is the paper's delivery contract. This analysis uses its own bootstrap seed, so an interval can differ in the last digit from Tables 5 and 16, where Ministral at 1024 tokens has [6.50, 14.17].

Setting	E1: EOS + final box	E2: final box, no EOS	E3: last box anywhere	E4: permissive
Qwen 2048, development	+7.81 [4.43, 11.72]	+7.81 [4.43, 11.72]	+8.07 [4.69, 11.98]	+7.29 [3.91, 10.94]
Qwen 2048, first rerun	+7.41 [4.88, 10.10]	+7.41 [4.88, 10.10]	+7.58 [5.05, 10.27]	+6.23 [3.87, 8.75]
Qwen 1024, second rerun	+9.33 [6.50, 12.33]	+9.17 [6.33, 12.33]	+9.33 [6.50, 12.50]	+8.67 [6.00, 11.67]
Qwen 2048, second rerun	+7.33 [5.00, 10.00]	+7.33 [5.00, 10.00]	+7.33 [5.00, 10.00]	+6.67 [4.33, 9.33]
Qwen 4096, second rerun	+4.50 [2.50, 6.83]	+4.50 [2.50, 6.83]	+4.83 [2.67, 7.33]	+4.50 [2.33, 7.00]
Ministral 1024, second rerun	+10.17 [6.67, 14.17]	+10.17 [6.67, 14.00]	+11.83 [8.00, 16.00]	+10.67 [7.00, 14.67]
Ministral 2048, second rerun	+6.83 [4.17, 9.83]	+6.83 [4.17, 9.83]	+7.00 [4.33, 10.00]	+6.00 [3.50, 8.83]

Answer origin. For every run active at the cue we take the text D and L share before it, which is identical in both arms by construction, so the audit cannot depend on the arm. From that text we extract candidate answers with their positions, keeping repetitions: every inline or display math segment of up to 200 characters and the right-hand side after its last "=", every complete box, and stated-answer phrases such as "answer is", "answer seems to be" and "final answer:". The correct value is *detected* if any candidate is equivalent to the reference under the paper's checker, with 10 seconds per check; ten checks in two runs timed out, and these two runs, neither rescued, count as not detected; every such run also contains incorrect candidates, and a run without a detection may still contain the value in a form the scan misses. The *latest* candidate is the one that ends last in the text. As a negative-control diagnostic, the same candidates are checked against the reference of the next question of the same setting whose reference is not equivalent to the true one. Table 18 crosses detection with the paired transition under E1 for the active runs D does not deliver correctly. The interval for the pooled difference in rescue rates, $+0.54$ [0.43, 0.63], resamples questions jointly across the settings that share them, within the development, first-rerun and second-rerun question sets, with 20000 draws; over the five second-rerun settings alone it is $+0.52$ [0.38, 0.64], 134/190 versus 100/530. Detection is a checker flag: in one rescued first-rerun run on a number-theory question the checker reads the divisibility statement $3 \mid 8$ as a bitwise expression equal to the reference 11. Excluding every candidate that contains one of the relation symbols $|$, $<$, $>$, $\leq$, $\geq$ or $\neq$ reclassifies only this run: 188/261 versus 121/654, $+0.54$ [0.43, 0.63], rescue origin 26/162/121. The first-occurrence scan, used for the frozen prediction below, orders candidates by first occurrence and keeps at most 60; its pooled rates are 179/244 versus 130/671.

Table 18: Answer origin among cue-active runs that D does not deliver correctly. Detected and not detected, as rescued by L over runs: whether an automated match finds the correct value among the candidates written before the cue. Placebo: runs flagged when a non-equivalent reference of another question is used. Rescue origin: counts of rescues in which the correct value is the latest candidate before the cue, is detected earlier but is not the latest, and is not detected.

Setting	Runs	Detected	Not detected	Placebo	Rescue origin
Qwen 2048, development	59	27/36	3/23	6	1/26/3
Qwen 2048, first rerun	136	28/36	17/100	19	3/25/17
Qwen 1024, second rerun	219	24/37	34/182	7	7/17/34
Qwen 2048, second rerun	110	35/45	10/65	5	4/31/10
Qwen 4096, second rerun	55	23/31	4/24	5	0/23/4
Ministral 1024, second rerun	216	28/42	34/174	5	8/20/34
Ministral 2048, second rerun	120	24/35	18/85	2	3/21/18
All	915	189/262	120/653	49	26/163/120

Falsifiable predictions for a registered follow-up. From these records we derived quantitative predictions for a separately registered follow-up that re-continues the archived second-rerun D streams with the same cue at other answer reserves r of 64, 256 and 512 tokens and with other wordings, and applies L to GPQA-Diamond. They were written down, with failure criteria, before any of its generations. They come from a simple exposure model computed from the records above, with gains proportional to the D runs active at the earlier cue that D does not deliver correctly and harms proportional to those it does; the model, its inputs, including a summary of the development matrix's earlier-cue arm, and the failure criteria are archived with the registration, and the predictions are reproducible from these records. The most consequential prediction concerns the budget decision: moving the cue from $r = 128$ to 256 changes the net gain by at most 8 runs of 600 per setting, whereas moving it to $r = 512$ lowers the net gain by about 11 runs, 1.8–1.9 points, in both 2048-token settings, so the smallest reserve that fits the answer contract should be chosen. For GPQA-Diamond the conversion among active runs is predicted to lie in [0.20, 0.60] with a harm rate of at most 0.15 among active runs D delivers correctly, and for every cue variant the rescue rate is predicted to be at least 0.25 higher when the first-occurrence scan finds the correct value before that variant's cue. Appendices E and F report the follow-up and the transfer, and Appendix Table 37 evaluates every prediction.

E Registered Follow-up: Answer Reserve and Answer-Field Cue

Design. A follow-up was registered in writing after the analyses of Appendix D and before any of its generations; its analysis script and the quantitative predictions below were frozen at the same time. It re-continues the archived D streams of three completed second-rerun settings, Qwen3.5-9B at 2048 and 4096 tokens and Ministral-3-8B at 2048, each with 200 questions and three streams. For each variant, every D stream still active at the variant's cue position $t = B - r - k$, with answer reserve r and k cue tokens, is continued from its own archived token prefix, with the cue tokens forced at t and native sampling with the same model, sampler and penalty scope until EOS or the total cap B; forced tokens count against B, and streams that end before t keep D's output, so every contrast uses all 600 runs. Variants: the registered cue at reserves $r = 64$, 128, 256 and 512; the box-opening cue of Table 34 at $r = 128$; and N0, the matched no-cue control that re-continues each stream from the registered cue position through the same path without any forced token. Two further wordings were registered as confirmatory comparisons and are listed in Table 35. Endpoint, scorer and inference are those of the second rerun, with question means over three streams and 20000 question-cluster bootstrap draws and sign flips under a fixed seed. The registration's confirmatory family H6 compares each alternative wording with N0 in the Qwen 2048-token setting, with Holm correction over three; all other contrasts are descriptive. The same registration's transfer to GPQA-Diamond is reported in Appendix F.

Execution. All 2521 continued runs completed under the registration's rules; no run was regenerated because of its outcome. The frozen analysis was then run once. Every continued run shares D's token

identifiers up to its cue position, carries the forced tokens exactly at $[t, t + k)$ and stays within the cap, and every D stream not continued had ended before the cue.

Table 19: Registered follow-up: continuations of the archived D streams under the delivery contract, 600 runs per row. Cue: position of the first forced token. Active: runs continued. Contrast: difference in points with 95% question-cluster interval against the matched no-cue control N0, and for N0 itself against the archived D. G/H: gains and harms in the same contrast. Tokens: generated tokens relative to the archived D, forced tokens included. The box-opening cue on Qwen 2048 is the confirmatory H6a of Table 35.

Variant	Cue	Active	Correct	Contrast [95%]	G/H	EOS	Tokens
Qwen, 2048: D 463, one-pass L 507 correct of 600							
N0, no cue, r=128	1916	121	459	D: −0.67 [−1.33, −0.17]	1/5	487	1.001
registered cue, r=64	1980	113	506	N0: +7.83 [5.33, 10.50]	48/1	594	0.990
registered cue, r=128	1916	121	505	N0: +7.67 [5.17, 10.33]	47/1	590	0.979
registered cue, r=256	1788	129	502	N0: +7.17 [4.83, 9.67]	43/0	591	0.954
registered cue, r=512	1532	152	484	N0: +4.17 [2.33, 6.17]	29/4	592	0.898
box-opening cue, r=128	1913	121	503	N0: +7.33 [4.83, 10.00]	45/1	592	0.978
Qwen, 4096: D 509, one-pass L 536 correct of 600							
N0, no cue, r=128	3964	55	510	D: +0.17 [0.00, 0.50]	1/0	546	1.000
registered cue, r=64	4028	55	535	N0: +4.17 [2.33, 6.33]	25/0	591	0.997
registered cue, r=128	3964	55	534	N0: +4.00 [2.17, 6.17]	24/0	594	0.992
registered cue, r=256	3836	59	532	N0: +3.67 [1.83, 5.83]	23/1	586	0.986
registered cue, r=512	3580	61	533	N0: +3.83 [2.00, 6.00]	23/0	592	0.966
box-opening cue, r=128	3961	55	537	N0: +4.50 [2.50, 6.83]	27/0	592	0.992
Ministral, 2048: D 425, one-pass L 466 correct of 600							
N0, no cue, r=128	1916	129	424	D: −0.17 [−0.83, 0.33]	1/2	481	1.000
registered cue, r=64	1980	124	466	N0: +7.00 [4.17, 10.17]	42/0	581	0.990
registered cue, r=128	1916	129	468	N0: +7.33 [4.50, 10.50]	44/0	586	0.976
registered cue, r=256	1788	140	460	N0: +6.00 [3.33, 9.00]	38/2	579	0.950
registered cue, r=512	1532	161	453	N0: +4.83 [2.17, 7.67]	33/4	581	0.888
box-opening cue, r=128	1912	129	454	N0: +5.00 [2.33, 8.00]	34/4	569	0.978

Calibration of the continuation path. N0 differs from the archived D by at most 0.67 points, four net runs of 600, in every setting, and the registered cue replayed at $r = 128$ differs from the one-pass L by 0.33 points, two runs of 600, in each setting. Re-continuing an archived prefix therefore delivers correct answers at rates within the registered tolerance of both arms, although the sampled continuations themselves differ, and the contrasts against N0 compare the forced tokens with a continuation resampled through the same path. Against the archived D, the registered cue at $r = 128$ gains $+7.00$ [4.67, 9.67], $+4.17$ [2.33, 6.33] and $+7.17$ [4.33, 10.33] points in the three settings.

Frozen predictions and their outcomes. Appendix Table 37 lists every frozen prediction, including those of Appendix D fixed before this follow-up, with its registered failure criterion and outcome. The quantitative model predicts net correct deliveries over D as

$$c_s \, \rho_r \, F_r - h_r \, K_r,$$

where F_r and K_r count the D streams active at the earlier cue that D does not and does deliver correctly, $c_s = G_{128}/F_{128}$ is the setting's one-pass rescue probability at $r = 128$ among the active runs D does not deliver correctly, 0.409, 0.491 and 0.350, a gross rate unlike the net conversion of Section 7; $\rho_r = 1$ up to $r = 256$ and $\rho_{512} = 0.885$, the ratio of the development matrix's rescue probabilities at the 512-token and 128-token reserves; and h_r is a harm rate, 0.081 up to $r = 128$ and 0.231 from $r = 256$, all computed from archived records only. Table 20 compares its cells with the outcomes.

Direct comparisons between cue wordings. After the frozen analysis, a descriptive comparison of the four wordings at the 128-token reserve was registered and run on the same cells with its own

Table 20: Frozen prediction and observed net correct deliveries over D, in runs of 600, of the registered cue at answer reserve r, continued from the archived D streams. Tolerance per predicted cell: $\pm \max(12, 0.3p)$ runs for a prediction of p runs; the $r = 128$ column is the replayed registered cue.

Setting	$r = 64$	$r = 128$	$r = 256$	$r = 512$
Qwen3.5-9B, 2048	44.3 / 43	44.1 / 42	41.3 / 39	32.5 / 21
Qwen3.5-9B, 4096	27.0 / 26	27.0 / 25	26.8 / 23	23.2 / 24
Ministral-3-8B, 2048	41.7 / 41	41.3 / 43	39.1 / 35	30.4 / 28

bootstrap seed and no decision rule; Table 21 reports it. The two wordings that name the final-answer field, the registered cue and the box-opening cue, exceed the two prose wordings of Table 34, a Time's-Up-style deadline and a plain stop instruction, in both Qwen settings, with every interval excluding zero; on Ministral the registered cue exceeds both prose wordings and the box-opening cue. The prose wordings are adapted to our reserve and total-cap accounting and are not re-executions of the protocols that introduced similar cues.

Table 21: Paired differences between cue wordings at the 128-token reserve, continued from the archived D streams under the delivery contract with 600 runs per wording, in points with 95% question-cluster intervals from 20000 draws; each cell gives the difference, its interval and the gains/harms of the first wording against the second. Wordings of Table 34: R, registered cue; BO, box-opening cue; DL, Time's-Up-style deadline; ST, plain stop instruction. Correct: counts for R, BO, DL and ST.

	Qwen, 2048	Qwen, 4096	Ministral, 2048
Correct, R/BO/DL/ST	505/503/462/460	534/537/511/509	468/454/457/448
R−DL	+7.17 [4.67, 9.83], 46/3	+3.83 [2.00, 6.00], 23/0	+1.83 [0.83, 3.00], 11/0
R−ST	+7.50 [5.00, 10.17], 47/2	+4.17 [2.33, 6.33], 25/0	+3.33 [1.50, 5.50], 21/1
BO−DL	+6.83 [4.33, 9.67], 45/4	+4.33 [2.33, 6.67], 26/0	−0.50 [−2.33, 1.17], 8/11
BO−ST	+7.17 [4.67, 10.00], 44/1	+4.67 [2.67, 7.00], 28/0	+1.00 [−1.00, 3.33], 15/9
BO−R	−0.33 [−1.83, 1.33], 9/11	+0.50 [−1.00, 1.83], 7/4	−2.33 [−4.33, −0.67], 2/16
ST−DL	−0.33 [−1.67, 1.00], 7/9	−0.33 [−0.83, 0.00], 0/2	−1.50 [−3.50, 0.33], 6/15

Scorer audit of the remaining comparisons of the reruns and the follow-up. The command-preserving comparator of Appendix B.3, applied so far to the development matrix, the first rerun and the second rerun's Qwen 2048-token arms, was then run as a registered descriptive check on every other MATH comparison in Table 22. It confirms 5048 of the 5090 positive verdicts. Of the 42 others, 41 are equivalent forms outside its grammar, namely column vectors written with fraction commands against a reference with slashes and a word answer written in a capitalized text wrapper; one continued run writes a column vector as a row vector, so removing it changes the affected contrast by 0.17 points. Crediting the delivered answers it finds equivalent but the paper scorer rejects leaves every L−D and every contrast of Table 19 at or above its reported value. GPQA-Diamond answers are single option letters and need no equivalence check. Table 28 in Appendix G audits the third study's continuations and numeric-budget generations.

Table 22: Command-preserving audit of the remaining comparisons under the delivery contract. Confirmed: positive verdicts of the paper scorer that the comparator confirms. Equivalent: delivered answers the paper scorer rejects but the comparator accepts. Credited contrast: with those answers credited, from 20000 question-cluster draws with its own seed; for the follow-up, the box-opening cue against N0.

Experiment	Arms	Confirmed	Equivalent	Credited contrast
Ministral-3-8B, 2048, H3	D, L, P	1299/1313	29	L−D +6.83 [4.00, 9.83], 42/1
Qwen3.5-9B, 4096, H4	D, L	1039/1045	29	L−D +4.67 [2.67, 7.00], 28/0
Qwen3.5-9B, 1024	D, L	787/788	10	L−D +10.00 [7.00, 13.17], 61/1
Ministral-3-8B, 1024	D, L	740/743	17	L−D +10.67 [7.00, 14.67], 65/1
P on the first rerun's questions	P	438/438	8	—
Follow-up, Qwen 2048	8 variants	291/292	15	+7.67 [5.17, 10.50], 47/1
Follow-up, Qwen 4096	8 variants	137/137	8	+4.67 [2.67, 7.00], 28/0
Follow-up, Ministral 2048	8 variants	317/334	14	+5.17 [2.50, 8.17], 35/4
Total		5048/5090	130	

F Registered Transfer to GPQA-Diamond

Design. The transfer was registered in writing together with the follow-up of Appendix E and before any of its generations, with its analysis script and the prediction P7 frozen at the same time. It uses all 198 GPQA-Diamond items (Rein et al., 2024); question order and option order are fixed by a seeded pseudorandom ordering, never by content or outcome. Qwen3.5-9B answers in direct mode with the paper's direct sampler, temperature 1.0, top-p 1.0, top-k 40 and presence penalty 2.0 on generated tokens, with the same three sampling seeds, batches of eight and rotated arm order. The system prompt is the math prompt with the solver described as a science solver for a multiple-choice problem and the answer placeholder changed to a letter; the user message is the stem followed by the options labelled A to D. D generates to EOS or the total cap of 2048 tokens; L forces the unchanged cue at generated position $2048 - 128 - 4 = 1916$, with forced tokens counted against the cap. The endpoint is the paper's delivery contract with a letter answer: EOS, one complete final box with no substantive trailing text, and a boxed value naming exactly one option letter equal to the gold letter, alone or in parentheses or a text wrapper and optionally followed by the option text. H5, L−D > 0, forms its own family and uses the inference of the second rerun, with question means over three streams and 20000 question-cluster bootstrap draws and sign flips. Three secondary read-outs were fixed in advance: any valid letter delivered, the final box with the gold letter without the EOS requirement, and a permissive final choice that takes the letter of the last complete box anywhere, else the last stated answer letter.

Execution and result. All 75 units, 25 question batches in three streams, completed, and every unit passed the D/L prefix identity and the total cap. The frozen analysis was run once on the complete set. Table 23 gives the counts. The registered mechanism statements hold: the 219 runs that end before 1916 tokens are token-identical in D and L, every change lies among the 375 active runs, and the gain of 120 runs is far below the active share. Of the 359 capped D outputs, 357 contain no final box. Of the 359 active runs in which D delivers no valid letter, L delivers a letter in 284, 120 correct and 164 incorrect, and none in 75; D delivers 14 correct letters in the active set and L keeps all 14. The permissive read-out, which ignores EOS and the final-field position, rises from 216 to 360, so the gain is not an artefact of the delivery contract. P7 predicted a conversion in [0.20, 0.60] with a harm rate of at most 0.15 among active runs D delivers correctly; the conversion is $120/375 = 0.32$ with no such harm.

Table 23: Registered transfer to GPQA-Diamond with Qwen3.5-9B, cap 2048, 198 questions and three streams. Correct: delivery contract with a letter answer. Active: D runs without EOS at the cue position 1916. H5: L−D $= +20.20$ points [16.50, 24.07]; 120 gains, no harm; 86 questions improve, none worsens; sign-flip $p = 5.0 \times 10^{-5}$.

Outcome, 594 runs	D	L	By domain	Runs	Active	D	L
Correct delivery	213	333	Physics	258	111	146	186
EOS	235	564	Chemistry	279	237	44	113
Capped	359	30	Biology	57	27	23	34
Any valid letter delivered	234	517	*375 active runs*				
Final box correct, EOS not required	213	333	Correct			14	134
Permissive final choice correct	216	360	Valid, incorrect			2	165
Generated tokens	1018880	977495	No valid delivery			359	76

G Registered Direct Comparisons, Reserve Controls and Cue Components

Design. A third study was registered in writing before any of its generations, with its analysis script and the held-out predictions below frozen at the same time. It has two parts. The first continues the archived D streams through the path of Appendix E, with the same model, sampler and penalty scope, forced tokens counted against the cap, and D's output kept for streams that end before a variant's cue, so every contrast uses all runs; it covers the three MATH settings and GPQA-Diamond. Its variants are the concluding message of Sun et al. (2025) as listed in Table 34, 14 tokens for both tokenizers, inserted at their trigger $B - 25$ under our total-cap accounting as a budget-charged variant in which message and answer share the last 25 tokens, and again at the registered trigger $B - 132$; the matched no-cue control N0 at the positions of the 64-, 256- and 512-token reserves; three component cues at the registered trigger, a box opener alone, a deadline alone and a field label that names neither an answer nor a stop, each four or five tokens long like the registered cue; and, on GPQA-Diamond, the registered cue at the other reserves with N0 at all four. Their prompt notice is not added, because it would change every token before the cue. Their original protocol allows 25 generated answer tokens after inserting the message, whereas the two placements we evaluate leave 11 and 118 answer tokens under full charging, so the reported contrasts compare budget allocations as well as cue text. The second part is a numeric-budget prompt Pn in the style of Han et al. (2025), generated from scratch on the second rerun's questions with the sampler, seeds and batching of P, using D's system prompt followed by the budget statement of Table 34. Like P, Pn receives the same cap and its extra prompt tokens are not charged. Endpoint, scorer and inference are those of Appendix E, with a fixed generator seed. The registered confirmatory family H7, with Holm correction over two members for Qwen at 2048 tokens, compares the one-pass L with Pn under H7a and the replayed registered cue with the budget-charged conclusion-message variant at $B - 25$ under H7b; all other contrasts are descriptive.

Execution. All 33 continuation configurations and the 150 Pn units ran to completion under the registration's rules; no run was regenerated because of its outcome. Every continued run is built from D's own token sequence up to its cue position, carries the forced tokens at $[t, t + k)$ and stays within the cap, and every D stream not continued had ended before the cue. The frozen analysis was run once on the complete set; Tables 24–27 report it, and Table 36 the confirmatory family H7.

The prior concluding message. At $B - 25$ the message leaves 11 tokens after its 14 forced tokens. It raises correct delivery over D in every setting, and the registered cue exceeds it with Qwen at 2048 and 4096 tokens and with Ministral. Scored with our final-field read-out without the EOS requirement, a narrower sensitivity than the prior protocol's own answer extraction, the message recovers a few more runs, shown in the column of final boxes without EOS, and the registered cue still exceeds it with Qwen at 2048 tokens. Moved to the registered trigger, the message gains $+2.17$ and $+3.17$ points more with Qwen at 2048 and 4096 tokens, so the answer room left at $B - 25$ accounts for much of the difference there; with Ministral the registered cue exceeds the message at either position by about 2 points, as Table 24 shows.

The numeric-budget prompt. Pn ends with EOS in 472 of 600 Qwen runs and 490 of 600 Ministral runs and delivers 450 and 437 correct answers, against 463 and 425 for D and 456 and 422 for P. Stating the numeric budget changes the whole generation: Qwen generates 9% more tokens than D and Ministral 9% fewer. The one-pass L exceeds Pn by $+9.50$ [6.17, 13.00] with Qwen under H7a and by $+4.83$ [0.83, 8.83] with Ministral in Table 25, while changing only the runs still active at its cue.

Components. At the registered trigger, with four or five forced tokens each, the box opener alone gains over N0 in all three MATH settings and the deadline alone gains with Ministral in Table 24. The registered cue exceeds each of the three components in all three settings of Table 4.

Answer reserve. N0 at every reserve delivers within 0.67 points of D on MATH and within 1.35 on GPQA-Diamond. Because N0 resamples the continuation of D's own prefix, it matches D's delivery rate approximately rather than token for token; its contrasts with D and their intervals are in Tables 19 and 24, and the reserve curves of Table 19 persist against it. Table 26 measures each reserve in the two ways of Section 4: against the stream stopped at the cue, the 512-token reserve has the largest gain in all three settings; against the stream continued to the cap, the 64-, 128- and 256-token reserves each gain more in both 2048-token settings. The 128-token reserve's advantage over 512 falls mostly on runs that both cues reach, as Proposition 3 separates.

Held-out reserve prediction. On GPQA-Diamond the registered cue delivers 338, 333, 315 and 309 correct answers at reserves of 64, 128, 256 and 512 tokens, where the 128 cell is the one-pass L. Table 27 compares the frozen predictions. The exposure model with its rescue rate re-estimated from the one-pass GPQA-Diamond cell has the lowest mean absolute error, 6.8 runs, and orders the three new cells correctly. The two model forms choose the 128-token reserve, within five correct deliveries of the 64-token cell at $+0.84$ [-2.19, 3.87] points.

Scorer audit of the third study. The command-preserving comparator of Appendix B.3 confirms 12205 of the 12295 positive MATH verdicts of the third study in Table 28. Of the 90 others, 88 are equivalent forms outside its grammar, namely column vectors written with fraction commands against a reference with slashes and a word answer in a capitalized text wrapper; the remaining 2 are numeric-budget outputs that the paper scorer accepts through its numeric tolerance, the rounded decimal 0.377 for $193/512$ and the nearby fraction $53/95$ for $82/147$. With the comparator's verdicts in place of the scorer's, removing unconfirmed positives and crediting equivalent rejected answers, every MATH contrast of Table 4 moves by at most 0.50 points; L$-$Pn becomes $+10.00$ [6.67, 13.50] with Qwen and $+4.83$ [1.00, 8.83] with Ministral.

The reserve under other answer read-outs. Table 29 repeats the comparison of the 128- and 512-token reserves under the delivery endpoint E1, the final field without EOS E2 and the last box anywhere E3, which on GPQA-Diamond is the permissive last-choice read-out. On GPQA-Diamond the 128-token reserve keeps its advantage under the permissive read-out, 360 against 339 correct, and in both 2048-token MATH settings the reserve ranking is the same under all three read-outs.

Table 24: Third registered study: continuations of the archived D streams under the delivery contract, with 600 runs per row and 594 on GPQA-Diamond. Cue: position of the first forced token, or for N0 the position at which the stream is re-continued. Active: runs continued. Contrast: against the matched no-cue control N0 at the registered trigger for cued variants at that trigger, against the archived D for the budget-charged Sun variants and N0 rows, and against N0 at the same position for the GPQA-Diamond reserve rows. Box, no EOS: correct final box without the EOS requirement. Tokens: generated tokens relative to the archived D, forced tokens included.

Variant	Cue	Active	Correct	Box, no EOS	Contrast [95%]	G/H	Tokens
Qwen, 2048: D 463, one-pass L 507 correct							
Sun message, $B-25$	2023	109	484	492	D: $+3.50$ [1.67, 5.67]	22/1	1.000
Sun message, $B-132$	1916	121	497	497	D: $+5.67$ [3.33, 8.17]	36/2	0.981
box opener	1916	121	494	494	N0: $+5.83$ [3.50, 8.33]	37/2	0.984
deadline	1916	121	466	466	N0: $+1.17$ [-0.17, 2.50]	12/5	0.996
neutral label	1916	121	456	456	N0: -0.50 [-1.50, 0.50]	3/6	1.001
N0, r=64	1980	113	461	462	D: -0.33 [-0.83, 0.00]	0/2	1.000
N0, r=256	1788	129	459	459	D: -0.67 [-1.67, 0.33]	3/7	1.001
N0, r=512	1532	152	461	462	D: -0.33 [-2.00, 1.33]	9/11	1.003
Ministral, 2048: D 425, one-pass L 466 correct							
Sun message, $B-25$	2023	121	455	458	D: $+5.00$ [2.50, 7.83]	31/1	1.000
Sun message, $B-132$	1916	129	456	456	D: $+5.17$ [2.83, 7.83]	32/1	0.983
box opener	1916	129	439	439	N0: $+2.50$ [0.83, 4.50]	19/4	0.991
deadline	1916	129	440	441	N0: $+2.67$ [1.00, 4.67]	18/2	0.997
neutral label	1916	129	422	422	N0: -0.33 [-1.17, 0.50]	2/4	1.001
N0, r=64	1980	124	425	425	D: 0.00 [0.00, 0.00]	0/0	1.000
N0, r=256	1788	140	426	426	D: $+0.17$ [-0.50, 0.83]	3/2	1.000
N0, r=512	1532	161	429	429	D: $+0.67$ [-0.50, 1.83]	8/4	0.999
Qwen, 4096: D 509, one-pass L 536 correct							
Sun message, $B-25$	4071	55	520	526	D: $+1.83$ [0.83, 3.00]	11/0	1.000
Sun message, $B-132$	3964	55	539	539	D: $+5.00$ [2.83, 7.50]	30/0	0.993
box opener	3964	55	523	523	N0: $+2.17$ [1.00, 3.50]	13/0	0.996
deadline	3964	55	511	511	N0: $+0.17$ [-0.33, 0.83]	2/1	0.999
neutral label	3964	55	509	509	N0: -0.17 [-0.50, 0.00]	0/1	1.000
N0, r=64	4028	55	510	510	D: $+0.17$ [0.00, 0.50]	1/0	1.000
N0, r=256	3836	59	513	513	D: $+0.67$ [0.00, 1.67]	4/0	1.000
N0, r=512	3580	61	513	513	D: $+0.67$ [0.17, 1.33]	4/0	0.999
GPQA-Diamond, Qwen, 2048: D 213, one-pass L 333 correct							
Sun message, $B-25$	2023	363	323	324	D: $+18.52$ [14.81, 22.39]	111/1	0.999
Sun message, $B-132$	1916	375	325	325	D: $+18.86$ [15.32, 22.56]	113/1	0.964
registered cue, r=64	1980	370	338	338	N0: $+21.21$ [17.34, 25.25]	127/1	0.981
registered cue, r=256	1788	385	315	315	N0: $+18.52$ [14.98, 22.22]	111/1	0.917
registered cue, r=512	1532	423	309	309	N0: $+17.17$ [13.30, 21.21]	110/8	0.818
N0, r=64	1980	370	212	212	D: -0.17 [-1.01, 0.67]	3/4	1.000
N0, r=128	1916	375	213	213	D: 0.00 [-1.18, 1.18]	5/5	1.000
N0, r=256	1788	385	205	205	D: -1.35 [-2.53, -0.34]	1/9	1.001
N0, r=512	1532	423	207	208	D: -1.01 [-2.69, 0.51]	6/12	1.000

Table 25: Third registered study: the numeric-budget system prompt Pn against the archived D, the one-pass L and the budget prompt P without a token count, on the second rerun's questions with 600 runs each, under the delivery contract with question-cluster intervals. Output: mean generated tokens.

	Qwen, 2048	Ministral, 2048
Pn correct	450	437
Pn EOS	472	490
Pn output	1062.8	787.0
L$-$Pn [95%], G/H	$+9.50$ [6.17, 13.00], 70/13	$+4.83$ [0.83, 8.83], 55/26
D$-$Pn [95%], G/H	$+2.17$ [-0.83, 5.17], 41/28	-2.00 [-5.83, 1.83], 31/43
P$-$Pn [95%], G/H	$+1.00$ [-1.67, 3.67], 38/32	-2.50 [-6.33, 1.17], 31/46

Table 26: The registered cue at answer reserve r measured against the same stream stopped at the cue, where D is correct only if delivered by the cue position as in Eq. (2), and against the archived D continued to the same cap as in Eq. (3); net correct deliveries in runs of 600. All cells are continuations of the archived D streams, and the $r = 128$ column is the replayed registered cue. Last column: Proposition 3 for $r = 512$ against $r = 128$, split into runs that end between the two positions and runs both cues reach.

Setting	Against	r=64	r=128	r=256	r=512	r=512 vs 128
Qwen, 2048	stopped at cue	+47	+53	+57	+59	
	continued D	+43	+42	+39	+21	−21 = −2 on 31 runs −19 shared
Ministral, 2048	stopped at cue	+45	+52	+52	+64	
	continued D	+41	+43	+35	+28	−15 = −2 on 32 runs −13 shared
Qwen, 4096	stopped at cue	+26	+25	+26	+29	
	continued D	+26	+25	+23	+24	−1 = 0 on 6 runs −1 shared

Table 27: Held-out reserve prediction on GPQA-Diamond, frozen before the continuations, as net correct deliveries of the registered cue over the archived D in runs of 594. M-transfer: the exposure model with every parameter calibrated on MATH; M-self: rescue rate calibrated on the GPQA one-pass cell at $r = 128$; R-const: the $r = 128$ net gain at every reserve; R-active: that gain scaled by the active count. Chosen: the reserve each predictor ranks first among 64, 128, 256 and 512, with ties going to the smaller reserve, and the correct deliveries observed there. MAE: mean absolute error of the forecasts, in runs.

	r=64	r=256	r=512	MAE	Chosen, correct
Observed	+125	+102	+96		64 (338)
M-transfer	146.5	142.1	118.2	27.9	128 (333)
M-self	119.7	114.5	93.5	6.8	128 (333)
R-const	120.0	120.0	120.0	15.7	64 (338)
R-active	118.4	123.2	135.4	22.4	512 (309)

Table 28: Command-preserving comparator audit of every positive MATH verdict of the third study, covering the eight continuation variants and the numeric-budget prompt Pn. Confirmed: positive verdicts of the paper scorer that the comparator confirms. Equiv. rejected: delivered answers the paper scorer rejects although the comparator finds them equivalent.

Setting	Output group	Confirmed	Equiv. rejected
Qwen, 2048	continuations	3753/3778	81
Qwen, 2048	Pn	448/450	8
Ministral, 2048	continuations	3458/3492	76
Ministral, 2048	Pn	432/437	9
Qwen, 4096	continuations	4114/4138	114
Total		12205/12295	288

Table 29: Correct deliveries of the registered cue at answer reserves 64–512 under three answer read-outs, E1 the delivery endpoint, E2 the final field without EOS and E3 the last box anywhere or, on GPQA-Diamond, the permissive last-choice read-out, with the paired contrast $R_{128}-R_{512}$ in points and 95% question-cluster intervals from 20000 draws. The 128-token cell is the replayed cue on MATH and the one-pass L on GPQA-Diamond. Runs per cell: 600 on MATH and 594 on GPQA-Diamond.

Setting	Read-out	D	$r = 64$	128	256	512	$R_{128}-R_{512}$
Qwen, 2048	E1	463	506	505	502	484	+3.50 [1.83, 5.33]
	E2	463	506	505	502	484	+3.50 [1.83, 5.33]
	E3	463	507	506	503	484	+3.67 [2.00, 5.67]
Ministral, 2048	E1	425	466	468	460	453	+2.50 [0.83, 4.33]
	E2	425	466	468	460	453	+2.50 [0.83, 4.33]
	E3	428	469	471	464	459	+2.00 [0.50, 3.67]
Qwen, 4096	E1	509	535	534	532	533	+0.17 [−0.83, 1.17]
	E2	509	535	534	532	533	+0.17 [−0.83, 1.17]
	E3	510	538	537	538	536	+0.17 [−0.67, 1.00]
GPQA-Diamond, Qwen, 2048	E1	213	338	333	315	309	+4.04 [1.01, 7.07]
	E2	213	338	333	315	309	+4.04 [1.01, 7.07]
	E3	216	358	360	344	339	+3.54 [0.51, 6.57]

H Fourth Registered Study: Prospective Reserve Choice and GPQA-Diamond with Ministral

Design. A fourth study was registered in writing before any of its generations, with its analysis and prediction scripts frozen at the same time. Stage 1 runs the unchanged policy on all 198 GPQA-Diamond questions with Ministral-3-8B, using the question order, options and prompt of Appendix F, the registered Ministral sampler, cap 2048, the cue at token 1916 and three streams, with D and L sharing one sampled stream per unit. Its confirmatory hypothesis H8, L−D in its own family, and a frozen prediction P10 of the conversion were fixed with it. Stage 2 continues the new D streams through the path of Appendix E with the registered cue at reserves of 64, 256 and 512 tokens and N0 at the 128- and 512-token positions. Before any output was read, an amendment added a reserve-selection family with Holm correction over three members. Each member compares the one-pass L, whose reserve is 128 tokens, with the cue at a 512-token reserve: H9 on GPQA-Diamond with Ministral, and H10a and H10b on the second rerun's questions at a 1024-token cap with Qwen and with Ministral, whose archived D streams were continued with the cue at 64, 256 and 512 tokens and N0 at 512. In these three operating conditions no reserve outcome had been generated before; the questions were known, since the 1024-token D and L arms were archived and GPQA-Diamond carried the Qwen transfer. The numerical forecasts P13 of the four predictors of Table 27 were fixed from archived records before any 1024-token continuation; for GPQA-Diamond with Ministral, the forecasts P11 were computed from the stage-1 D and L records after stage 1 and before stage 2 by a script fixed in advance. A second amendment, also before any output was read, fixed a completeness gate, under which a member can be confirmed only if no continuation configuration or eligible record is missing from the records its analysis reads and every cell passed the prefix and cap checks; it also kept Holm over the three planned members regardless of which were executed, and added the paired contrasts of the reserves the predictors chose. Endpoint, scorers and inference are those of Appendix G.

Execution. All 75 stage-1 units and all 13 continuation configurations ran once to completion. Every stage-1 unit passed the prefix-identity and cap checks, every continued run stays within the cap, the completeness gate lists no missing record, a check of the expected question, stream and arm keys finds all 1188, 1200 and 1200 baseline D and L records of the three conditions present exactly once, a validation step that withholds any confirmation unless that check is complete leaves every label of the gate unchanged, and each frozen analysis was run once on the complete set. Tables 30–33 give the results, and Tables 36 and 37 list the registered tests and predictions.

GPQA-Diamond with Ministral. D reaches the cap in 110 of 594 runs, 108 of them without any box, and 112 runs are active at the cue. Under H8, L gains $+4.55$ points [2.69, 6.57] with $p = 5.0 \times 10^{-5}$, 27 rescues and no harm; 22 questions improve and none worsens. L generates 1.9% fewer tokens, and the 482 runs that end before the cue are token-identical in D and L. The conversion of .24 lies in the frozen interval [0.15, 0.50] of P10. The final-box read-out without EOS and the permissive read-out give $+4.55$ and $+4.88$ points.

Reserve choice at a 1024-token cap. A 512-token reserve places the cue at token 508, where 392 Qwen and 298 Ministral streams are still active, against 245 and 232 at the registered position. The one-pass L exceeds it by $+14.33$ points [10.67, 18.33] with Qwen under H10a, with 93 gains and 7 harms, and by $+10.33$ [6.50, 14.33] with Ministral under H10b, with 67 gains and 5 harms; both have Holm-adjusted $p = 1.5 \times 10^{-4}$. Under the final-box read-out without EOS and the permissive read-out of Table 31, the differences are $+14.33$ and $+14.50$ points with Qwen and $+10.33$ and $+12.00$ with Ministral. N0 at the 512-token position delivers within 0.67 points of D in both settings, so the continuation path can account for only a small part of these differences. The long reserve saves generated tokens, 29% and 26% fewer than L. Proposition 3 splits the 86-run Qwen advantage of the 128-token reserve into 40 runs among the 147 runs that end between the two positions, 141 of which D delivers correctly, and 46 among the runs both cues reach; for Ministral the parts are 12 among 66 runs and 50 shared, as Table 33 shows. At a 1024-token cap both parts of the exchange therefore favour the short reserve.

Scorer audit of the fourth study. The registered counts use the fixed scorer, whose numeric tolerance is $10^{-4}\max(1, |g|)$ on MATH. For example, one Ministral no-cue continuation at the

512-token position delivers 0.377 for a probability whose reference is $193/512 = 0.376953125$; the difference lies within the tolerance, so the scorer accepts it, and this run belongs to N0, not to either arm of H10b. A post hoc audit applies the command-preserving comparator of Appendix B.3 to all 7200 stored MATH answer records of this study, including rejected deliveries. It confirms 4525 of 4539 positive verdicts and finds 87 further equivalent pairs. Counting only comparator-confirmed deliveries, the 128-minus-512 reserve differences are $+14.67$ [10.83, 18.67] points for Qwen and $+10.17$ [6.50, 14.17] for Ministral, from 20000 question-cluster draws, so the reserve ordering holds under this sensitivity analysis. A stricter partial exact numeric comparator that keeps unsupported expressions unresolved and counts only its confirmed equalities gives $+10.67$ [7.33, 14.17] and $+8.50$ [5.00, 12.50] points. Both audits read the stored answer fields; the registered operational endpoint remains the primary result.

Frozen reserve predictions. In Table 32, M-transfer, M-self and R-const chose the 64-token reserve in all three conditions and R-active chose 512. The 64-token reserve delivered the most correct answers in all three conditions; at 1024 tokens it exceeds the 512-token reserve by $+16.00$ points [12.17, 20.00] with Qwen and $+12.00$ [8.00, 16.33] with Ministral in Table 31. R-const chose the same reserve, so these contrasts validate the short-reserve decision itself rather than one selector over the constant rule. M-self, whose rescue rate is re-estimated from each condition's one-pass cell, has the lowest forecast error in each condition, with mean absolute errors of 2.5, 22.0 and 22.5 runs.

The comparison again decides the ranking. Against the stream stopped at the cue, the 512-token reserve has the largest gain with Qwen at 1024 tokens, $+137$ runs, and on GPQA-Diamond with Ministral, $+37$; against the stream continued to the same cap the 64-token reserve is first in all three conditions, as Table 33 shows.

Table 30: Fourth registered study: cells under the delivery contract, with 594 runs per row on GPQA-Diamond and 600 at the 1024-token cap. L is the one-pass arm with the registered 128-token reserve; the other rows continue the D streams from their own token prefix. Cue: position of the first forced token, or for N0 the position where the stream is re-continued. Active: runs continued. Box, no EOS: correct final box without the EOS requirement. Permissive: last choice on GPQA-Diamond, last complete box anywhere on MATH. Tokens: generated tokens relative to D, forced tokens included.

Cell	Cue	Active	Correct	Box, no EOS	Permissive	EOS	Tokens
GPQA-Diamond, Ministral, 2048							
D			241	241	246	484	1.000
L, r=128	1916	112	268	268	275	572	0.981
cue, r=64	1980	110	271	271	278	572	0.991
cue, r=256	1788	119	264	264	269	568	0.961
cue, r=512	1532	135	266	266	275	578	0.907
N0, r=128	1916	112	241	241	246	485	1.000
N0, r=512	1532	135	240	240	246	481	1.002
MATH, Qwen, 1024							
D			366	367	367	382	1.000
L, r=128	892	245	422	422	423	590	0.939
cue, r=64	956	232	432	432	433	592	0.973
cue, r=256	764	280	398	398	398	588	0.867
cue, r=512	508	392	336	336	336	594	0.669
N0, r=512	508	392	362	363	363	381	1.004
MATH, Ministral, 1024							
D			341	341	343	386	1.000
L, r=128	892	232	402	402	414	556	0.941
cue, r=64	956	221	412	412	420	569	0.973
cue, r=256	764	247	384	384	391	566	0.864
cue, r=512	508	298	340	340	342	565	0.697
N0, r=512	508	298	344	344	346	386	1.003

Table 31: Fourth registered study: paired contrasts in points under the three read-outs of Table 30, with gains/harms; contrasts use question means over three streams and 20000 question-cluster bootstrap draws. The Delivery values of H8 and of $R_{128}-R_{512}$ at 1024 tokens are the registered analyses; the other intervals come from the frozen descriptive contrast script, which uses its own generator seed.

Contrast	Delivery [95%]	G/H	Box, no EOS [95%]	Permissive [95%]
GPQA-Diamond, Ministral, 2048				
L−D, H8	+4.55 [2.69, 6.57]	27/0	+4.55 [2.69, 6.57]	+4.88 [2.69, 7.24]
$N0_{128}-D$	0.00 [0.00, 0.00]	0/0		
$N0_{512}-D$	−0.17 [−1.01, 0.51]	2/3		
MATH, Qwen, 1024				
$R_{128}-R_{512}$, H10a	+14.33 [10.67, 18.33]	93/7	+14.33 [10.67, 18.17]	+14.50 [10.83, 18.50]
$R_{64}-R_{512}$	+16.00 [12.17, 20.00]	104/8	+16.00 [12.17, 20.00]	+16.17 [12.33, 20.17]
$R_{64}-R_{128}$	+1.67 [0.00, 3.50]	21/11	+1.67 [0.00, 3.50]	+1.67 [0.00, 3.50]
$N0_{512}-D$	−0.67 [−2.50, 1.17]	13/17		
MATH, Ministral, 1024				
$R_{128}-R_{512}$, H10b	+10.33 [6.50, 14.33]	67/5	+10.33 [6.50, 14.50]	+12.00 [8.17, 16.17]
$R_{64}-R_{512}$	+12.00 [8.00, 16.33]	78/6	+12.00 [7.83, 16.33]	+13.00 [9.00, 17.33]
$R_{64}-R_{128}$	+1.67 [−0.67, 4.17]	20/10	+1.67 [−0.67, 4.17]	+1.00 [−1.00, 3.17]
$N0_{512}-D$	+0.50 [−0.33, 1.33]	6/3		

Table 32: Frozen reserve forecasts in the three new operating conditions, in net correct deliveries of the cue over D out of 594 runs on GPQA-Diamond and 600 at 1024 tokens. Predictors as in Table 27: M-transfer calibrated on the same model at 2048 tokens on MATH, M-self with its rescue rate from the condition's one-pass cell. Chosen: the reserve each predictor ranks first among 64, 128, 256 and 512, with the correct deliveries observed there. MAE: mean absolute error of the forecasts, in runs.

	r=64	r=256	r=512	MAE	Chosen, correct
GPQA-Diamond, Ministral, 2048; one-pass L: +27					
Observed	+30	+23	+25		64 (271)
M-transfer	38.5	38.2	35.3	11.3	64 (271)
M-self	27.0	26.4	23.9	2.5	64 (271)
R-const	27.0	27.0	27.0	3.0	64 (271)
R-active	26.5	28.7	32.5	5.6	512 (266)
MATH, Qwen, 1024; one-pass L: +56					
Observed	+66	+32	−30		64 (432)
M-transfer	88.1	76.2	42.9	46.4	64 (432)
M-self	56.7	44.4	14.2	22.0	64 (432)
R-const	56.0	56.0	56.0	40.0	64 (432)
R-active	53.0	64.0	89.6	54.9	512 (336)
MATH, Ministral, 1024; one-pass L: +61					
Observed	+71	+43	−1		64 (412)
M-transfer	74.3	70.2	56.1	29.2	64 (412)
M-self	61.0	56.4	43.2	22.5	64 (412)
R-const	61.0	61.0	61.0	30.0	64 (412)
R-active	58.1	64.9	78.4	38.1	512 (340)

Table 33: Fourth registered study: the cue at reserve r against the stream stopped at the cue, Eq. (2), and against D continued to the same cap, Eq. (3), in net correct deliveries; the $r = 128$ column is the one-pass L. Last column: Proposition 3 for $r = 512$ against $r = 128$, split into the runs that end between the two positions and the runs both cues reach.

Setting	Against	r=64	r=128	r=256	r=512	r=512 vs 128
GPQA-Diamond, Ministral, 2048	stopped at cue	+30	+29	+29	+37	
	continued D	+30	+27	+23	+25	−2 = −1 on 23 runs −1 shared
MATH, Qwen, 1024	stopped at cue	+80	+82	+92	+137	
	continued D	+66	+56	+32	−30	−86 = −40 on 147 runs −46 shared
MATH, Ministral, 1024	stopped at cue	+78	+77	+71	+66	
	continued D	+71	+61	+43	−1	−62 = −12 on 66 runs −50 shared

I Implementation Details

I.1 Experimental design and implementation

Figure 3 summarizes the paired design. The unit of the development matrix is one of 128 exposed MATH questions, selected from the previously inspected 897-item complement by a pseudorandom ordering of question identifiers fixed before any output. No gold value enters selection or the policy. An earlier greedy run used these same questions, so the sampling matrix is a development study. All 48 batches of eight questions, across three seeds, contain all five arms, 1920 outcomes in all. No cap, failed delivery, subject or seed is excluded.

The model is Qwen3.5-9B. The model-card revision selected before the experiment has two inconsistent non-thinking recommendation blocks. Before any output, the protocol chose the final Best Practices block: D temperature 1, top-p 1, top-k 40, presence penalty 2; T temperature 1, top-p .95, top-k 20, presence penalty 1.5. Both use repetition penalty 1 and min-p 0. The presence penalty applies once to each token type already generated and excludes prompt tokens; its cost is inside the measured time. The card's typical 32768-token output allowance and its 81920-token allowance for difficult benchmarks are far above this deliberately tight 2048-token experiment.

Sampling streams are fixed by replicate, batch and phase. D, L and C share the D seed; T and F share the T seed; the restarted C stage has a separate fixed phase seed. Arm order rotates across batches. Generation is native; D, L, T and F each use a single call, with no copied cache or re-tokenized text. C alone discards the direct prefix and pays a second prompt prefill.

D uses its whole output allowance. L injects four tokens after 1916 only if no EOS has occurred, leaving 128 answer tokens. T has no intervention; F injects the six-token close-and-answer suffix after 1024 unless thinking has already closed. C runs D to 1024, returns an EOS-completed D answer, and otherwise restarts T with at most 1024 tokens; the restarted stage forces after 890, charging six instruction tokens and reserving 128. The forced suffixes are fixed token sequences of the model's tokenizer, and natural EOS and the thinking close are its dedicated special tokens; the supplementary archive lists their identifiers. The output allowance includes every suffix.

I.2 Records and independent rescoring

All 1920 stored development outcomes were independently regraded with the fixed scorer, and all stored token sequences were independently decoded with the same tokenizer; every text, suffix and endpoint matched. All 1152 prefix equalities and all 1920 budget checks pass. Independent rescoring uses the original tokenizer and the stored outputs and does not repeat inference. Model weights and implementation remain unchanged during data collection.

For the development matrix and both reruns, the supplementary archive contains the per-run outcome table, with the delivered answer, verdict, EOS, cap, cue position and token counts of every run, from which every count and paired contrast is recomputed, together with the questions, the official reference solutions and the scripts that regrade raw records. It also contains the raw records of the development matrix, both reruns, the secondary experiments, the follow-up and both GPQA-Diamond studies, with the complete generated token sequences and reproducible output text, endpoint flags, forced positions and per-batch costs, and the continuations of the third and fourth studies from their cue on.

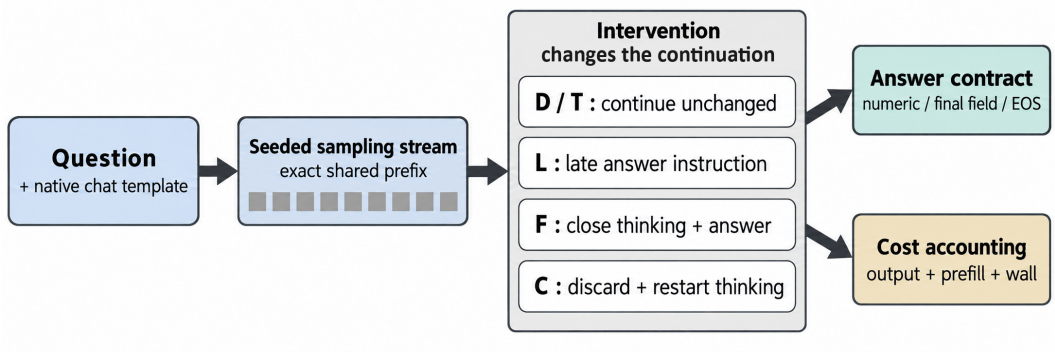

Figure 3: **Interventions, outcomes and costs are separate parts of the comparison.** A fixed question and seeded sampling stream produce an exact shared token prefix: D and L share the direct prefix, T and F share the thinking prefix, and C shares D's prefix only up to its restart decision; each intervention changes only what follows. Every arm is assessed under an explicit answer contract, while cost accounting includes all work it executes; prefix equality removes pre-intervention variation from each prefix-sharing contrast.

Table 34: Protocol card and forced texts. The development study and the first rerun use Qwen3.5-9B, total output cap $B = 2048$, batch size eight and three fixed sampling seeds. Arms D, L, T, F and C use the same system and user messages in the model's native chat template, and only the thinking flag differs; P and Pn append the instructions shown to D's system message. Every forced token counts toward B. D is told the required answer format, so L adds a reminder, not new information. Prompts and forced texts are verbatim, and a line break is shown as ↩. Backslashes, braces, angle brackets, dots and asterisks inside them are literal characters of the prompts; in F the thinking-close token is shown by its text, and the box-opening cue deliberately ends with an open brace that the model completes.

System message	`You are a careful math solver. Solve the problem step by step briefly. End with a single line: Final answer: \boxed{<answer>}.`
User message	The MATH problem statement, unchanged.
D	Thinking disabled; generate to EOS or B.
L	As D with the same seed; if no EOS by 1916, force the four tokens ↩`Final answer:` and continue natively; 128 tokens remain.
T	Thinking enabled; generate to EOS or B.
F	As T with the same seed; if thinking is still open at 1024, force the six tokens ↩`</think>`↩↩`Final answer:` and continue.
C	Run D to 1024; keep an EOS-completed answer, otherwise discard it and run T in the remaining 1024 tokens with F's suffix after 890; the second prefill and discarded tokens are charged.
P	D's system message followed by `Solve the problem within the available token budget. Keep the reasoning concise and reserve enough space to state one complete final answer in \boxed{...}. Do not end before giving that answer.`
Pn	D's system message followed by `Let's think step by step and use less than 1920 tokens, so that one complete final answer in \boxed{...} fits within the 2048-token output limit. Do not end before giving that answer.`
Box-opening cue	↩`Final answer: \boxed{`
Prose cues	Deadline: ↩↩`Time's up! I have to state my final answer now.`↩; stop instruction: ↩↩`I should stop here and give the final answer.`
Components	Box opener ↩`\boxed{`; deadline ↩`Time's up.`; neutral label ↩`Next step:`
Sun message	↩`**Time's Up!**`↩`Therefore, the final answer is:`, 14 tokens for both tokenizers
Endpoint	Natural EOS, thinking closed where applicable, one complete last balanced box with no substantive trailing text, accepted against the reference answer by the unchanged MATH scorer, including its numeric tolerance.
Unit	Question; the three streams are averaged before paired bootstrap with 20000 draws and sign-flip tests.

J Full Accounting Identities

All probabilities refer to a fixed question distribution, sampler and answer contract. L is the unconstrained natural total output length; the thinking-close position is a separate observable. Conditional accuracies belong to the subsets defined at their stated budget and need not stay constant as that budget changes.

J.1 Shared-prefix gain versus equal-total gain

Proposition 4. Shared-prefix extraction identity. *Suppose completed outputs at b_r are returned unchanged under both procedures. If an answer stage has correctness $\alpha_e(b_r, b_a)$ on the unfinished group, its gain over the original score at b_r is Eq. (2); its gain over continuation to $b_r + b_a$ is Eq. (3).*

Proof. Step 1: alternative score. The alternative score is

$$S = F_L(b_r)\alpha_c(b_r) + \big(1 - F_L(b_r)\big)\alpha_e(b_r, b_a).$$

Step 2: gain over stopping at b_r. Subtracting Eq. (1) at b_r cancels the completed group:

$$S - \mathrm{Acc}_{\mathrm{T}}(b_r) = \big(1 - F_L(b_r)\big)\big[\alpha_e(b_r, b_a) - \alpha_t(b_r)\big],$$

which is Eq. (2).

Step 3: gain over continuation to $b = b_r + b_a$. Subtract Eq. (1) at b, then add and subtract $F_L(b_r)\alpha_c(b)$ and $\big(1 - F_L(b_r)\big)\alpha_t(b)$:

$$\begin{aligned} S - \mathrm{Acc}_{\mathrm{T}}(b) &= F_L(b_r)\alpha_c(b_r) + \big(1 - F_L(b_r)\big)\alpha_e(b_r, b_a) \\ &\quad - F_L(b)\alpha_c(b) - \big(1 - F_L(b)\big)\alpha_t(b) \\ &= \big(1 - F_L(b_r)\big)\big[\alpha_e(b_r, b_a) - \alpha_t(b)\big] \\ &\quad + \big[F_L(b_r) - F_L(b)\big]\alpha_c(b) + \big[F_L(b) - F_L(b_r)\big]\alpha_t(b) \qquad \text{(i)} \\ &\quad + F_L(b_r)\big[\alpha_c(b_r) - \alpha_c(b)\big]. \end{aligned}$$

Step 4: collect terms. The two terms in row (i) combine into one product:

$$\begin{aligned} &\big[F_L(b_r) - F_L(b)\big]\alpha_c(b) + \big[F_L(b) - F_L(b_r)\big]\alpha_t(b) \\ &\quad = -\big[F_L(b) - F_L(b_r)\big]\big[\alpha_c(b) - \alpha_t(b)\big], \end{aligned}$$

which gives Eq. (3). □

The second line of Eq. (3) measures newly completed mass; the third measures changed completed-group accuracy. The paired design of the main text measures continued generation directly, so neither term has to be neglected.

J.2 Localization and the choice of reserve

Proof of Proposition 2. Step 1: outside A. If D has emitted EOS by t, L never inserts its cue and, by prefix identity, returns the same completed output, so the two policies have the same correctness there.

Step 2: decomposition. Total probability over A and its complement gives

$$\begin{aligned} \mathrm{Acc}_{\mathrm{L}} - \mathrm{Acc}_{\mathrm{D}} &= \Pr(A)\big[\Pr(\text{L correct} \mid A) - \Pr(\text{D correct} \mid A)\big] \\ &\quad + \Pr(A^c) \cdot 0, \end{aligned}$$

which is Eq. (4).

Step 3: bounds. Under an EOS-gated endpoint a capped run delivers nothing, so inside A D is correct only if it ends correctly in $(t, B]$. The lower bound follows from $\Pr(\text{L correct}, A) \geq 0$; the upper bound from $\Pr(\text{L correct}, A) \leq \Pr(A)$. □

Proof of Proposition 3. Outside A' both policies return D's completed output. On $A' \setminus A$ the stream ends by t, so L returns D's output while L$'$ acts on it. On A both act. Total probability over the three events gives the identity. □

The first two lines are the price of an earlier cue: on $A' \setminus A$ every run that D delivers correctly by t can only be lost or kept, so together they are at most $\Pr(A' \setminus A) - \Pr(A' \setminus A,\ \text{D correct})$ and at least $-\Pr(A' \setminus A,\ \text{D correct})$. The third line is the change in conversion on the runs both cues reach, which has no sign in general. Choosing a reserve therefore trades these terms at the fixed cap; Appendix Table 26 measures them, and Appendix G adds matched no-cue continuations at every reserve.

K Registered Confirmatory Tests and Frozen Predictions

This appendix lists every registered confirmatory test of the reruns, the registered follow-up, the registered transfer and the third and fourth registered studies in Tables 35 and 36, and every prediction frozen before the follow-up and the later studies in Table 37, each with the analysis and decision rule fixed in advance and its outcome.

Table 35: Registered confirmatory analyses of the two reruns on unused questions, the registered follow-up and the registered transfer, as specified in their written plans. Differences are in points, averaging each question over its three streams. Holm-adjusted p values are over the registered family H1–H4 and, separately, H6a–c, H7a–b and H9–H10b; H5 and H8 each form their own family. H6 is reported from the frozen analysis run when the follow-up was complete; the run on the complete set after the transfer finished, whose draws follow H5, gives the same estimates, counts and decisions, with H6a interval [4.83, 10.17].

Hypothesis	Registered analysis	Result	Decision rule
First rerun: L−D, Qwen3.5-9B, B=2048	numeric-identity comparator; separate-endpoint identification bounds, 20000-draw question-cluster envelope	bounds $[-10.33, +28.67]$, envelope $[-16.33, +34.50]$	envelope lower endpoint > 0
H1: L−D, Qwen3.5-9B, B=2048	delivery scorer; bootstrap interval, sign-flip p, Holm over H1–H4	+7.33 [5.00, 10.00]; $p = 5.0 \times 10^{-5}$; Holm $p = 2.0 \times 10^{-4}$	Holm $p < .05$, estimate > 0
H2: L−P, same setting, two-sided	as H1	+8.50 [5.67, 11.50]; $p = 5.0 \times 10^{-5}$; Holm $p = 2.0 \times 10^{-4}$	as H1
H3: L−D, Ministral-3-8B, B=2048	as H1	+6.83 [4.17, 9.83]; $p = 5.0 \times 10^{-5}$; Holm $p = 2.0 \times 10^{-4}$	as H1
H4: L−D, Qwen3.5-9B, B=4096	as H1	+4.50 [2.50, 6.83]; $p = 5.0 \times 10^{-5}$; Holm $p = 2.0 \times 10^{-4}$	as H1
Registered transfer, App. F: GPQA-Diamond, Qwen3.5-9B, B=2048			
H5: L−D	delivery scorer, letter answer; bootstrap interval, sign-flip p; own family	+20.20 [16.50, 24.07]; 120/0; $p = 5.0 \times 10^{-5}$	$p < .05$, estimate > 0
Registered follow-up, App. E: Qwen3.5-9B, B=2048, archived D streams re-continued, $r = 128$			
H6a: box-opening cue − N0	delivery scorer; bootstrap interval, sign-flip p, Holm over H6a–c	+7.33 [4.83, 10.00]; 45/1; $p = 5.0 \times 10^{-5}$; Holm $p = 1.5 \times 10^{-4}$	Holm $p < .05$, estimate > 0
H6b: Time's-Up-style deadline − N0	as H6a	+0.50 [−0.67, 1.67]; 10/7; $p = .58$; Holm $p = 1$	as H6a
H6c: plain stop instruction − N0	as H6a	+0.17 [−1.17, 1.50]; 8/7; $p = 1$; Holm $p = 1$	as H6a

Table 36: Registered confirmatory analyses of the third and fourth registered studies, continuing Table 35 with the same conventions; cue wordings as in Table 34.

Hypothesis	Registered analysis	Result	Decision rule
Third registered study, App. G: Qwen3.5-9B, B=2048, second rerun's questions			
H7a: one-pass L − Pn	delivery scorer; bootstrap interval, sign-flip p, Holm over H7a–b	$+9.50$ [6.17, 13.00]; 70/13; $p = 5.0 \times 10^{-5}$; Holm $p = 1.0 \times 10^{-4}$	Holm $p < .05$, estimate > 0
H7b: replayed registered cue − concluding message at $B-25$	as H7a	$+3.50$ [1.50, 5.67]; 26/5; $p = 1.1 \times 10^{-3}$; Holm $p = 1.1 \times 10^{-3}$	as H7a
Fourth registered study, App. H: Ministral-3-8B on GPQA-Diamond; reserve choice in three new conditions			
H8: L−D, GPQA-Diamond, Ministral-3-8B, B=2048	delivery scorer, letter answer; bootstrap interval, sign-flip p; own family; completeness gate	$+4.55$ [2.69, 6.57]; 27/0; $p = 5.0 \times 10^{-5}$	$p < .05$, estimate > 0, complete
H9: L − cue at r=512, same setting	delivery scorer; bootstrap interval, sign-flip p; Holm over H9, H10a, H10b; completeness gate	$+0.34$ [-0.84, 1.52]; 7/5; $p = .77$; Holm $p = .77$	Holm $p < .05$, estimate > 0, complete
H10a: L − cue at r=512, MATH, Qwen3.5-9B, B=1024	as H9	$+14.33$ [10.67, 18.33]; 93/7; $p = 5.0 \times 10^{-5}$; Holm $p = 1.5 \times 10^{-4}$	as H9
H10b: L − cue at r=512, MATH, Ministral-3-8B, B=1024	as H9	$+10.33$ [6.50, 14.33]; 67/5; $p = 5.0 \times 10^{-5}$; Holm $p = 1.5 \times 10^{-4}$	as H9

Table 37: Predictions frozen before the generations they concern, with their registered failure criteria and outcomes. Net gains are correct deliveries over the archived D, in runs of 600 or 594 on GPQA-Diamond.

	Prediction	Fails if	Outcome
P1	net gain at $r = 512$ below that at $r = 128$ in each setting	reversed in ≥ 2 settings	21 vs 42, 24 vs 25, 28 vs 43: holds in all three
P2	more harms at $r = 512$ than at $r = 128$	not so in ≥ 2 settings	6 vs 1, 0 vs 0, 3 vs 0: holds in two of three
P3	$\lvert \text{N0} - \text{D} \rvert < 2$ points in every setting	any setting ≥ 2	-0.67, $+0.17$, -0.17: holds
P4	$\lvert$replayed cue $-$ one-pass L$\rvert < 2$ points	any setting ≥ 2	-0.33, -0.33, $+0.33$: holds
P5	nine reserve cells of Table 20 within $\pm \max(12, 0.3p)$ runs of prediction p	≥ 3 cells outside	0 of 9 outside: holds
P6	$r = 256$ within 8 runs of $r = 128$; $r = 512$ below $r = 128$ in both 2048-token settings	$r = 512 \geq r = 128$ in both, or $r = 256$ more than 8 below in ≥ 2 settings	-3, -2, -8; -21 and -15: holds
P7	conditional net gain on the second benchmark in [0.20, 0.60]	outside, or harm rate > 0.15 with at least 10 active runs D delivers correctly; untestable if fewer than 2% of D runs are active	0.32, 120 net of 375 active; no harm among 14 active runs D delivers correctly: holds
P8	rescue rate higher by ≥ 0.25 when a correct value is found before the cue, for each of the seven cued variants	below 0.25 for ≥ 2 variants	0.43–0.58 for the five variants reported in Table 19; 0.20 and 0.12 for the two other registered wordings
P9	GPQA-Diamond net gains at $r = 64$, 256, 512 from the four frozen predictors of Table 27	descriptive; no failure criterion	MAE 27.9, 6.8, 15.7 and 22.4 for the MATH-calibrated model, the GPQA-calibrated rescue rate, the constant and the active count
P10	conditional net gain of L over D on GPQA-Diamond with Ministral-3-8B in [0.15, 0.50]	outside; untestable if fewer than 30 D runs are active	0.24, 27 net of 112 active: holds
P11	same setting: net gains at $r = 64$, 256, 512 from the four predictors of Table 32, computed after stage 1 and before stage 2; P11a: correct delivery at $r = 512$ below $r = 128$; P11b: the best reserve is 64 or 128	forecasts descriptive; P11a and P11b as stated	MAE 11.3, 2.5, 3.0, 5.6 in the order of Table 32; 266 vs 268: P11a holds; best 64: P11b holds
P12	in each new condition the stopped comparison ranks $r = 512$ first and the continued comparison ranks 64 or 128 first, with ties to the smaller reserve	per condition, as stated	GPQA-Diamond with Ministral and MATH with Qwen at 1024: holds; MATH with Ministral at 1024: the stopped comparison ranks 64 first in Table 33
P13	1024-token net gains at $r = 64$, 256, 512 from the four predictors of Table 32, fixed before the continuations	descriptive; no failure criterion	MAE 46.4, 22.0, 40.0, 54.9 for Qwen and 29.2, 22.5, 30.0, 38.1 for Ministral; three predictors choose 64, the observed best, and R-active 512

For P8, the rescue rate among cue-active runs that D does not deliver correctly, when the first-occurrence scan finds a correct value before the variant's cue versus when it does not, is 75/106 versus 35/178, 74/97 versus 37/188, 59/87 versus 42/203 and 56/80 versus 26/215 for the registered cue at $r = 64$, 128, 256 and 512, and 62/96 versus 40/189 for the box-opening cue, differences of 0.43–0.58; three checks timed out in each variant and count as not found.