Answer Interfaces Change the Measured Cost of Early Stopping

Anonymous authors
Paper under double-blind review

Abstract

Early-exit methods are commonly compared by reasoning tokens saved and accuracy lost, as if the loss belonged to the stopping rule. Yet a stop returns a reasoning prefix; a terminal interface must still turn it into a scored answer. We hold rule-selected prefixes fixed and cross the interfaces that read them. On Qwen3.5-9B, the same Answer Convergence (AC) stops lose 35.5 accuracy points under a staged cascade inside the open thinking block and 7.0 after model-specific closure. Stopping changes how often the answer stage sees an open, unresolved prefix, while requests, prefills and closed generation respond differently to that state. At matched mean reasoning length and a 1K answer allowance, AC trails a fixed-token cut by 9.5 points under the cascade but leads by 8.0 after closure. In matched-prefix controls, two closed readouts agree within a pre-specified margin at 2K. The dependence recurs across caps, tasks and checkpoints, while the response to closure and prefill differs across model families. We therefore report complete systems under their own procedures and compare rules conditionally on a specified answer interface.

1 Introduction

Early-exit papers summarize a system with two numbers: reasoning tokens saved and accuracy lost. This convention treats the loss as a property of the stopping rule. But a stop returns a prefix, not a submission. An answer procedure still decides whether to close the thinking block, ask for an answer or prefill its opening, how long to generate, and which text to score. If those choices fail differently at rule-selected endpoints and full endpoints, the reported cost—and even a comparison between methods—can change without changing a single retained token.

We isolate this dependence by placing adapted EAT and Answer Convergence (AC) stops on shared trajectories and giving the generating interfaces exactly the same retained prefix. On AMC23, the same AC stops cost 35.5 points under a staged cascade inside the open thinking block and 7.0 after model-specific closure. The interface effect differs between full endpoints and early stops, so the relevant quantity is the change in the full-to-stop gap, not accuracy under either interface alone. The consequence is practical: at the same mean reasoning length and a 1K answer allowance, AC trails a fixed cut by 9.5 points under the cascade but leads by 8.0 after closure.

Stopping methods do use different terminal procedures: some return a short trial answer, others close the thinking block and resume generation, and others prefill or append an answer cue (Liu & Wang, 2025; Wang et al., 2025; Song et al., 2025; Fu et al., 2025; Muennighoff et al., 2025; Wang et al., 2026; Yang et al., 2025b). Most of the nine published procedures audited in Appendix C.2 close the block or prefill the answer's opening; the practical ambiguity is broader than the open-block failure. Methods still differ in answer mode, reader and allowance, and EAT generates until the end of the sequence without fixing an answer allowance for a capped comparison. Answer templates, forced extraction and regeneration are known to change scores (Wang & Zhu, 2026; Jo et al., 2025; Dong et al., 2026), and reasoning models often disregard instructions placed inside their reasoning (Kwon et al., 2025). We measure the non-additive effect of these choices: how an interface changes accuracy differently at rule-selected stops and full endpoints, and how that difference changes policy comparisons.

We call everything that turns a retained prefix into a submitted string the *terminal interface*: block state, answer mode, instruction, decoding, allowance and submitted text; the probe that chooses

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

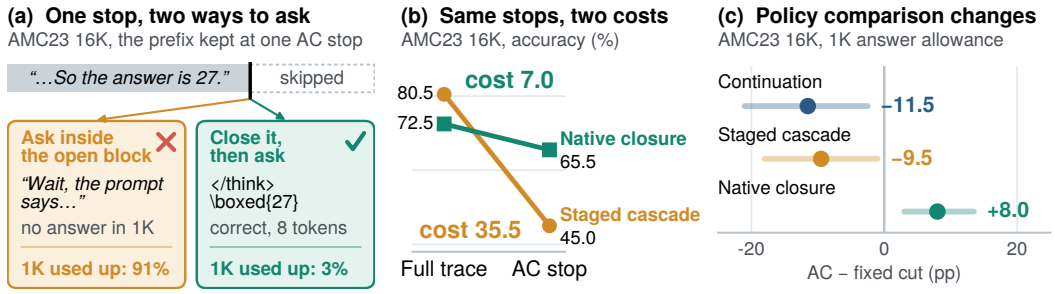

Figure 1: **The terminal interface changes both a stop's measured cost and a rule comparison.** (a) One retained prefix, answered by a request inside the open thinking block or after closing it (Figure 3); the bottom line is the share of the 200 AMC23 trajectories under AC on which that request uses up its 1K allowance. (b) On AMC23 the staged cascade is more accurate than native closure on full endpoints but much less accurate on the same AC stop prefixes; the stop costs are 35.5 and 7.0 points. (c) AC minus a fixed token cut matched to its mean reasoning length, on AMC23 with a 1K answer allowance. The comparison changes from -9.5 points under the cascade to $+8.0$ after closure (bars: pointwise paired 95% intervals; Table 1).

the stop belongs to the rule. In the generating-interface crossing, each interface receives identical retained tokens, so its difference at a stop arises after that stop; the historical stored-trial readout at some horizon-cut full endpoints uses the last checked boundary instead (Appendix A.3). We find:

- **The same stop has different measured costs.** In the initial, jointly inferred crossing, every detected interaction involves an early exit; a larger prospective experiment reproduces the effect where most full traces end on their own (Section 3).
- **Evaluation conclusions change.** At a 1K answer allowance, AC reverses against a fixed cut when the interface changes; comparisons with a smaller source cap and adapted EAT also depend on the answer procedure (Section 4).
- **Stopping changes exposure; the interface sets the response.** Early stops present the answer stage with more open, unresolved prefixes. For Qwen3.5, an answer request inside the open block often elicits further reasoning, whereas closure or an answer prefill elicits a submission; other model families respond differently (Section 5).

Section 6 tests the scope, including a non-mathematical task and a model family on which the mechanism differs. Section 8 turns the results into a reporting checklist.

2 Setup

Objects. A source trajectory ends when it self-terminates or reaches its allowance; we call that retained text the *full endpoint*, which may therefore be horizon-cut. An early-exit policy P instead returns a prefix ending at a *stop point*; it *fires* when that point precedes the full endpoint. An interface F turns either retained text into the string that is scored. *Reasoning tokens* precede the first end-of-thinking token. The crossing's cap B covers retained source tokens and a final 1K answer allowance; probe and trial-answer costs are reported separately. For the mechanism analysis, a prefix is *eligible* when its thinking block is open and the task reader finds no answer in its retained text. A submission is *valid* when the same reader finds a legal final answer (full definitions: Appendix A).

Measured cost. Write $\mathrm{Acc}(P, F)$ for accuracy under source policy P and interface F with everything else fixed. The measured accuracy cost and the policy–interface interaction are

$$C(P, F) = \mathrm{Acc}(\mathrm{Full}, F) - \mathrm{Acc}(P, F), \qquad I(P; F_a, F_b) = C(P, F_b) - C(P, F_a),$$
$$I(P; F_a, F_b) = \big[\mathrm{Acc}(P, F_a) - \mathrm{Acc}(P, F_b)\big] - \big[\mathrm{Acc}(\mathrm{Full}, F_a) - \mathrm{Acc}(\mathrm{Full}, F_b)\big]. \tag{1}$$

Both are in percentage points; positive I means stopping costs less under F_a. I is the change of an interface's effect between the stops and the full traces; it is not the value of the discarded reasoning, because an interface still computes after the stop. For the strict pair below, whose two conditions

make the same call or none on every ineligible source, the contrast τ_P (strict closed minus strict open on the eligible sources of P) and the eligible share q_P give $I = q_P\tau_P - q_{\text{Full}}\tau_{\text{Full}}$: early stops leave more eligible sources (exposure), and those prefixes may respond differently from eligible horizon cuts.

Source policies. The crossing starts from one sampled trajectory per problem and seed: Qwen3.5-9B (Qwen Team, 2026) on AMC23 (40 problems), AIME25 (30) and SMT25 (53) (MathArena, 2026), five seeds each, at caps of 16K, 16K and 32K. *Adapted EAT-v1* (Wang et al., 2025) checks entropy stability after a hypothetical end-of-thinking token at blank-line boundaries, with the 20-check warm-up of arXiv v1 (v3 replaces it with bias correction; Appendix A.1); the original paper's experiments add a "Final answer: " probe prefix, a variant tested in the appendix. Its designated threshold in each setting is the one whose reasoning-token saving is closest to 15%, chosen without using accuracy. *Adapted Answer Convergence* (AC) (Liu & Wang, 2025) stops after ten consecutive identical greedy trial answers at the same boundaries, which are coarser than the original sentence boundaries (Appendix A.3). *s1-style forcing* (Muennighoff et al., 2025) suppresses two attempted endings by appending "Wait"; it serves as a control because it does not create an early open prefix. Appendix A.1 gives the exact rules, threshold sweep and probe costs.

Terminal interfaces. The initial crossing uses three 1K interfaces. *Open continuation* keeps sampling inside the thinking block and serves as a diagnostic control. The *staged cascade* checks for a candidate and conditionally rescues, requests, corrects and finalizes it, also inside the block. *Native closure* appends the model family's closure package—for Qwen3.5, a budget sentence followed by the end-of-thinking token—then answers freely and greedily. Its 25 injected tokens count toward the 1K. Mechanism controls use the same strict final-answer request either inside the block (*strict open*) or after closure (*strict closed*), with a full 1K answer allowance in both conditions. The *stored boxed trial* is AC's greedy answer of at most 48 tokens after closure and the prefilled opening "The final answer is \boxed{". Self-terminated traces are returned unchanged, and every interface uses the same scorer. Appendix A.1 gives the exact prompts.

What was fixed in advance. The initial, internally time-stamped family was fixed before generation: 27 interactions (three policies, three interface contrasts and three task–cap settings) in one simultaneous problem-cluster bootstrap band (20,000 draws, with five seeds kept together). These interfaces differ in decoding and instructions as well as block state, so this family establishes interface dependence; Section 5 locates the component with controls and follow-ups. We mark families fixed before their own runs and post hoc analyses separately (Appendix A). Most follow-ups use AMC23 16K, SMT25 32K and AMC23 32K; the complete crossing contains twelve settings.

3 The interface changes a stop's measured cost

On AMC23 the same AC stops cost 28.5 points more under the open-block cascade than after native closure (Figure 1). AC fires on 169 of 200 trajectories and saves 51.5% of reasoning tokens across the full set; roughly half of the full endpoints are themselves horizon-cut. Closure gains 20.5 points at AC's stops but loses 8.0 at the full endpoints, which is why the change in stop cost is larger than the stop-only gain. At the exact endpoint, the three closed readouts put AC's cost between 7.0 and 9.5 points, whereas open continuation and a strict request inside the block put it above 50 (Figure 2; Appendix A.3).

In the initial simultaneous band, all five detected interactions involve early exits. Five of 27 contrasts exclude zero in the common ± 16.0-point simultaneous band (two contrasts per policy–setting pair are linearly independent, giving 18 in total). They are the AC contrasts between native closure and the two open interfaces on AMC23 and SMT25, plus EAT's closure–continuation contrast on AMC23; no forcing-control or AIME25 contrast excludes zero. AIME25's full endpoints are usually open and unanswered already, so early stopping adds few such prefixes. Conversely, the effect remains in the cleaner AMC23 32K setting, where 181 of 200 full traces end naturally: its AC closure–continuation interaction is 52.5 points, the largest in the crossing (Table 4).

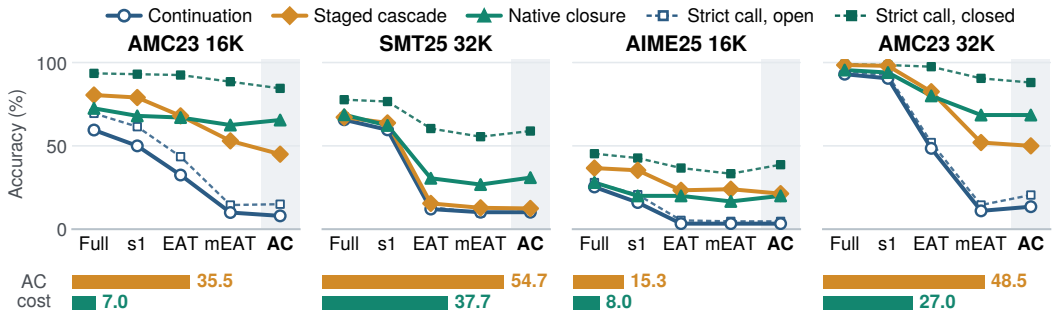

Figure 2: **The same stops cost far more when the answer is requested inside the open block than after closure.** Three primary settings and AMC23 32K. Continuation is a diagnostic control; policies appear in the same order in every panel; EAT is the designated threshold and mEAT the matched EAT of Section 4, the threshold 2^{-1} whose saving is closest to AC's. Dashed lines read the same stops with one strict final-answer call at 1K, inside the open block or after closure; the bars below each panel give AC's stop cost, full endpoint minus AC stop, under the staged cascade and under native closure. Intervals: Table 4.

4 The interface changes evaluation conclusions

Rule against a fixed cut of the same length. A fixed token cut matched to AC's mean reasoning length removes as many reasoning tokens but places its stops without reading the trace (a control fixed before its run; Appendix A.3). With a 1K answer allowance, AC is 9.5 points less accurate on AMC23 than this cut under the staged cascade and 8.0 points more accurate after native closure, a 17.5-point reversal; on SMT25 the difference moves 13.2 points in the same direction (Table 1). Under the stored trial, their point differences stay within four points and each interval includes zero. This is a comparison of complete policies at the same mean reasoning length, so it includes which trajectories are cut as well as where. On the same fired trajectories, AC's selected locations add 3.8 to 8.8 points over depth-matched random cuts under closed readouts (Section 5).

Table 1: **At the same mean reasoning length, the interface can reverse the complete-policy comparison.** AC minus a fixed token cut (pp) at a 1K answer allowance, with pointwise paired 95% intervals; n gives trajectories cut by each policy. The last column is the change from staged cascade to native closure.

Cell	n_{AC}	n_{cut}	Continuation	Cascade	Closure	Closure − cascade
AMC23 16K	169	177	−11.5 [−21.0, −2.5]	−9.5 [−18.0, −1.0]	+8.0 [3.0, 13.5]	+17.5 [7.5, 27.5]
SMT25 32K	194	211	−13.2 [−23.8, −3.4]	−16.2 [−27.2, −5.7]	−3.0 [−12.1, 5.7]	+13.2 [5.3, 21.5]
AIME25 16K	88	136	−6.7 [−16.0, 0.7]	−2.7 [−11.3, 6.7]	+3.3 [−0.7, 8.0]	+6.0 [−4.0, 15.3]

Procedure-specific readouts. Across ten Qwen3.5-9B settings, AC's 48-token stored trial leads matched EAT-v1's 1K native closure by 18 to 39 points. With EAT-v1's paper suffix and up to 8K answer tokens, no setting retains a statistically detected AC lead (Figure 9); the three primary settings show the same change at 2K. This comparison measures the complete adapted procedures, including their answer allowances. In the three primary settings, the matched EAT-v1 threshold is the loosest grid point and concentrates exits at the end of warm-up (Appendix A.1).

Rule against a smaller cap. On AMC23, AC under a 16K cap (retaining 5,972 source tokens on average) is 8.5 points more accurate than a plain 8K cap (6,946 tokens) under native closure ($[2.5, 15.0]$) but 9.5 points less accurate under the cascade ($[-17.0, -2.0]$); under the common strict closed readout the two differ by -1.0 points at 1K ($[-7.5, 5.0]$) and 0.0 at 2K, so no accuracy advantage is detected under those conditions. The serving regime then decides the cost: under native closure AC at 16K costs 0.89 times the tokens of the 8K cap with a persistent KV cache but 4.2 times as many if every trial re-reads its prefix (Table 17).

A prospective comparison on MATH-500. On 127 level-5 problems with five seeds and a 32K cap, AC minus matched EAT-v1 is $+28.2$ points under procedure-specific 1K readouts and -1.4 under the EAT-style 8K readout ($[-4.7, 1.6]$). The recorded common-2K outputs give 0.0 ($[-2.0, 2.0]$), but include requests incorrectly skipped when the entry reader accepted a placeholder. These outputs are retained as a diagnostic of the implemented pipeline; the corrected validity rates and repair sensitivity are reported in Appendix B.4.

A validity screen flags visible submission failures. A shared readout puts rules under the same terminal conditions only where it submits answers, so the checklist of Section 8 first screens it for validity. At 1K the screen accepts strict closed on AMC23 and flags SMT25 and AIME25 for a larger allowance, so that a difference such as the 5.3-point AC$-$EAT gap on AIME25 is not read as a difference between rules; a comparison that passes still keeps its readout conditions and residual invalid share (Appendix B.4).

5 Where the dependence comes from

Early stops leave unanswered prefixes, and only some interfaces turn them into submissions. For the strict pair, Equation 1 separates *exposure*—how often a policy leaves an open prefix with no readable answer (q)—from *response*, the closed-minus-open difference on such prefixes (τ). AC leaves about 90% of trajectories in this state in each control setting, versus 40.0%, 30.9% and 7.5% at the corresponding full endpoints. At AC's AMC23 stops, for example, continuation submits a valid answer on only 8% of trajectories and exhausts its allowance on 97%; native closure is valid on 66%. The descriptive validity term accounts for almost all of the closure–continuation interaction and most of the cascade–closure interaction (Table 2).

The paired transitions show what this descriptive validity term contains. Among 546 AC stops in the control settings, 368 move from invalid under strict open to correct under strict closed; at 356 of those stops, AC's stored 48-token trial is also correct. Thus a short readout of the retained prefix supports the answer in most repaired cases, although that trial also participates in selecting AC's stops. A wrong-prefix control is correct on at most 7.1% of stops and usually submits the answer belonging to the substituted problem (Table 12).

Exposure carries most of the strict-pair interaction: on AMC23 about 36 of 45.5 points come from the larger eligible share, and exposure is the larger component in nine of ten estimable settings. The interaction is small when the conditional interface response is near zero—under answer prefill and at DeepSeek's blank-line stops—and the earliest sentence-level AC stops have a larger exposure but no detected strict-pair interaction.

Table 2: **The interaction is concentrated in valid-submission rates.** A descriptive split of the AC-minus-endpoint interaction (pp) into a validity term (pointwise 95% interval) and the remainder from conditional correctness. Both contrasts point toward native closure, so positive means a lower measured cost under native closure. $*$: symbolic task, where the remainder also includes a small fallback-reader term.

	Native closure − continuation			Native closure − staged cascade		
Cell	Total	Validity	Conditional	Total	Validity	Conditional
AMC23 16K	$+44.5$	$+44.5$ [31.0, 57.5]	$+0.0$	$+28.5$	$+17.0$ [11.5, 22.5]	$+11.5$
SMT25 32K*	$+17.7$	$+16.9$ [9.7, 24.7]	$+0.8$	$+17.0$	$+16.6$ [9.5, 24.4]	$+0.4$
AIME25 16K	$+14.0$	$+14.0$ [3.3, 26.0]	$+0.0$	$+7.3$	$+5.7$ [-0.4, 12.4]	$+1.6$

The effect of closing the block depends on how the answer is elicited. Holding the strict instruction, greedy decoding and 1K allowance fixed, adding the end-of-thinking token raises accuracy by 47.6 to 74.4 points in the three control settings; the budget sentence stays within a few points of zero (Figure 8). Inside the open block, the output typically restarts with "Wait," re-checks the problem and runs out of room (Figure 3); changing the request's wording does not remove the pattern (Appendix B.2). When the answer's opening is prefilled instead, the six delimiter effects lie between -2.1 and $+0.6$ points. The response therefore depends on both prefix state and answer mode, rather

Retained prefix, ending: *"Position of meeting* $= 18(1.5) = 27$. *Since* $0 < 27 < 45$, *the point is between A and B. So the answer is 27."*

✗ **Open, requested (1K).** *"Wait, the prompt says 'Give the final answer as a signed integer in \boxed{N}.' Usually, distance is non-negative. Is there any ambiguity? Maybe the coordinate system? […]"* The output re-derives the meeting point in other coordinate systems until all 1,024 tokens are used and ends mid-equation (*"Position* $= 0 - 18$*"*): no valid answer, scored wrong.

✓ **Closed, requested (1K).** After the closure package: `\boxed{27}` (8 tokens), correct.

✓ **Open or closed, prefilled (48).** After *"The final answer is \boxed{"*: *"27}."* (4 tokens), correct in both.

Figure 3: **One stop, four readouts: only the request inside the open block fails to submit.** A stop of Answer Convergence on AMC23 16K (Qwen3.5-9B); quoted text is verbatim model output, elided at […]. In this cell the strict call inside the open block exhausts its allowance on 91% of the 200 trajectories under AC and after closure on 3% (Appendix B.2).

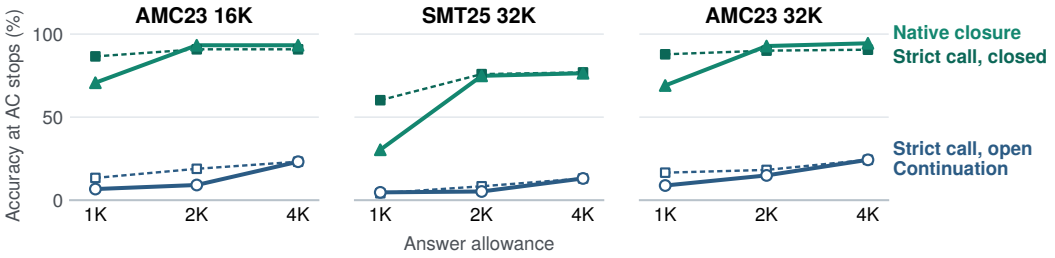

Figure 4: **More room helps free answers after closure but does not rescue the open request at AC stops.** Accuracy on identical AC stop prefixes as the answer allowance grows. Open-block interfaces remain below 25% at 4K, whereas native closure reaches the strict closed readout by 2K. This is a mechanism diagnostic: the added allowance may take total generation beyond the original cap. Other policies and empty traces are in Table 14.

than on closure alone (Table 13). Across the 432 cases in the original Qwen3.5-9B crossing cells where AC's stored trial is correct and a fresh native closure is not, the free answer has exhausted its 1K allowance (Appendix B).

Allowance affects free generation and strict requests differently. With every source fixed and the allowance raised to 4K (a mechanism diagnostic that exceeds the cap), a strict call inside the open block still exhausts it on about 90% of eligible AC stops and trails the same call after closure by 56.9 to 69.4 points, all excluding zero in the twelve-contrast band. In these controls, the strict-closed minus native-closure difference falls from $+24.0$ points at 1K to -0.6 at 2K (pooled; 90% interval $[-3.2, 1.9]$), equivalent within the ± 5 points fixed before the run, well below the 17.7–52.5-point interactions of the control cells (Figure 4; Table 14).

When the stop comes too early, closing the block helps no more than at the full trace. Checked at sentence ends, as in the original rule, AC stops at a median 3.3% of the reasoning, where its trial answer is right only 34.0% of the time; the stop still costs more under continuation than after closure, and a strict call after closure still beats one inside the block by 21.5 points ($[10.0, 33.5]$), but no more than the 24.0 points it adds at the full traces, so the strict pair's interaction is not detected (-2.5, $[-18.5, 13.0]$): at these stops the short trial is usually incorrect (Appendix A.3). With an empty trace, native closure solves 39.0% of AMC23 within 1K and 70.5% within 4K, whereas a strict call after closure stays at 16.0% at every allowance, consistent with a free answer that still reasons after closure rather than with a better readout.

The interaction extends beyond AC-selected locations. Depth-matched random cuts give closure-versus-continuation interactions of $+37.0$, $+12.2$ and $+46.3$ points in the control settings (means of five seeded shuffles), against AC's $+44.5$, $+17.7$ and $+52.5$. AC's selected locations add 3.8 to 8.8 points under the four closed readouts, in every shuffle, as expected of a rule that selects its stops with a closed trial answer (Appendix A.3). Nor does the gap come from interrupting fail-

ing traces: at AC's AMC23 stops on 92 traces that would have ended correctly on their own, native closure answers 95.7% correctly and continuation 12.0% (Table 5).

6 Scope: caps, tasks, checkpoints and source generation

Across the twelve crossing settings, native closure lowers the point estimate of stop cost relative to continuation by 23.7 points for AC and 18.0 for EAT on average, against 1.8 for s1-style forcing; detection in an individual setting depends on its family's band (Figure 5). Seven replications add tasks and checkpoints: all 14 early-exit estimates are positive, seven exclude zero in their own bands, and no forcing-control estimate does. Sources generated in one request give the same pattern, showing that the pattern persists without resetting the presence penalty between chunks; the larger MATH-500 experiment reproduces it where 92% of full traces end naturally (Table 7).

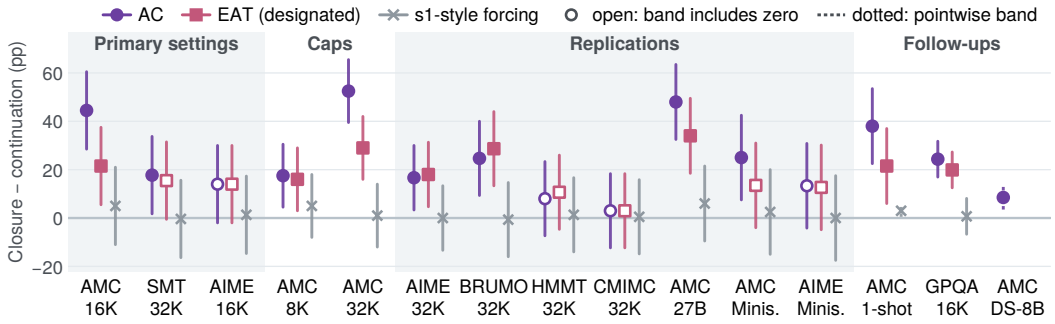

Figure 5: **Interface dependence recurs across settings, while its magnitude and mechanism vary.** Closure-versus-continuation interaction (pp) of AC, the designated EAT threshold and s1-style forcing, with the simultaneous band for each setting's family (dotted: pointwise); open markers include zero. Bands from different families are not one simultaneous family. Minis.: Ministral-3-8B; DS-8B: DeepSeek-R1-Distill-Llama-8B, AC only; 1-shot: sources generated in one request. Stop costs under both interfaces are in Figure 6.

The dependence has the same shape on a non-mathematical task. On GPQA-Diamond (Rein et al., 2023) (198 questions, 16K), AC's stops cost 31.9 points under continuation and 7.6 under native closure, and six of the eleven pre-specified contrasts exclude zero in their simultaneous band, none of them for s1-style forcing. The official reader takes the first parenthesized letter anywhere in the source and the generated answer, so most source endpoints already hold a readable answer and the stored trial costs nothing; AC's closure-versus-continuation interaction is $+24.3$ with the official reader (excluding zero in the family band), $+19.9$ with the answer region alone and $+12.9$ with its last letter (pointwise $[14.9, 25.1]$ and $[8.3, 17.7]$) (Appendix B).

The mechanism differs on a second model family. On DeepSeek-R1-Distill-Llama-8B, whose native closure is its end-of-thinking token alone, native closure still lowers AC's cost relative to continuation by 8.5 points (pointwise $[3.5, 13.5]$), but a strict call inside the open block already yields a valid answer on 65% of eligible AC stops, against 9% for Qwen3.5-9B on AMC23, and closing the block with the end-of-thinking token changes that call's accuracy by -0.6 points ($[-13.8, 12.7]$, simultaneous): which part of the interface matters differs between model families (Appendix B).

Sentence-boundary AC on a model from its original study. On DeepSeek-R1-Distill-Llama-8B, we apply AC's sentence-end stopping definition ($k = 10$) and cross our terminal interfaces. On AMC23 the rule saves 66% of reasoning tokens, and the stored 48-token trial has a 12.5-point lower stop cost than free closure ($[3.0, 23.0]$). On AIME24, free-closure accuracy falls from 40.0 to 16.7% and no interface interaction is detected (Table 16). Appendix B.5 specifies the differences between our requests and the released AC script.

7 Related work

Stopping and budget-control work decides when to interrupt reasoning (Muennighoff et al., 2025; Liu & Wang, 2025; Wang et al., 2025; Yang et al., 2025b; Fu et al., 2025); answer-formation work decides what a retained trace can submit (Jo et al., 2025; Liao et al., 2026; Xu et al., 2025; Dong et al., 2026). Three measurement studies are closest. Wang & Zhu (2026) find that free continuation recovers far more than a forced boxed answer from short fixed fractions of a trace; there continuation may reason to completion, here it must finish inside a 1K allowance, and the two studies differ in prefix depth, remaining allowance and block state, factors varied in our controls. Dong et al. (2026) compare early exits under one protocol and several serving regimes and find answer templates fragile; we hold the stops fixed and cross them with interfaces. Carlon et al. (2026) probe reasoning at matched horizons and report completion and the stopped-prefix answer separately. Early answering truncates a chain of thought to probe faithfulness (Lanham et al., 2023); its answer request is itself a terminal interface. Reasoning models often ignore instructions placed inside their reasoning (Kwon et al., 2025); our open-block failures are an instance in the answer stage, but a prefilled answer opening avoids them and DeepSeek-R1-Distill-Llama-8B follows the same request on 65% of eligible stops, so the failure depends on how the answer is elicited and on the model. Our paired crossing measures the non-additive interface effect between full endpoints and rule-selected stops, and directly tests whether two policies respond differently to the same interface change. Appendix C.2 audits the published procedures and positions other answer-formation work.

8 Discussion, reporting checklist and scope

The Qwen3.5 controls distinguish three regimes of a stopped prefix. At the earliest stops no tested interface reliably recovers a correct answer, and closing the block helps a strict request no more than at the full trace. At AC's blank-line stops a short prefilled request usually produces a correct answer, whereas a request inside the open block often fails to produce a valid submission within its allowance; closing the block or prefilling the answer's opening elicits one. At 2K, the two tested closed readouts satisfy the pre-specified equivalence margin on the pooled control estimand. Across these regimes early stopping sets the exposure, how many unresolved prefixes reach the interface, and the interface, with the prefix state, sets the response to them; model families differ in the response. The measured cost of a stop is therefore a property of the joint system—the rule, its probe, the terminal interface, the reader and the allowance—and not the value of the discarded reasoning. Systems can be compared as deployed; a common readout supports a rule comparison conditional on that readout, while an explicit crossing shows how the comparison changes across readouts.

Reporting checklist. An early-exit evaluation should

(i) name the rule and its probe, and the terminal interface: block state, answer mode (request or prefill), instruction, decoding and allowance, how the submitted string is built and which reader scores it;

(ii) report each complete system under its own answer procedure; for a comparison conditional on a shared readout, avoid reusing a compared rule's own probe and check that readout on the target model and task for validity on every stop type (a declared screening level, 95% here), prefix sensitivity, allowance sensitivity and cost; state the reader, candidate set and any selection performed on evaluation data, and report the range across qualifying interfaces. If none qualifies, report the complete systems and sensitivity to residual invalid submissions, without a shared-readout ranking;

(iii) report the correct, valid-wrong and invalid shares, at least two allowances, and the stop-event composition (self-terminated, horizon-cut, rule-selected);

(iv) report probes, replayed input and calls under a stated serving regime, and compare rules at matched source savings or along a threshold curve.

The 1K strict closed readout passes the screen on AMC23; the screen flags SMT25 and AIME25 for different terminal conditions. On Ministral, a third-party prefill passes on AMC23 and changes AC$-$EAT from -8.5 to $+7.5$, while strict requests terminate immediately. The MATH-500 audit identifies false acceptance of placeholders as a source of skipped requests (Appendix B.4). These ap-

plications illustrate the screen's role in detecting submission failures; accuracy comparisons remain conditional on the readout and reader.

Limitations. The main rules use adapted EAT-v1 and blank-line AC; their operating points define the policy comparisons. The fixed-cut reversal holds at a 1K answer allowance, while equivalence of the two closed readouts is assessed on the pooled control settings. Most main tasks have 30–53 independent problems, yielding a simultaneous half-width of about 16 points (8 on MATH-500). Token accounting represents stated serving regimes, rather than measured latency or energy. Model-specific request construction and run-time readers remain part of the evaluated interface.

Reproducibility Statement

Appendices A–C specify the crossing, its replications, controls and paired statistical families, including the decoding settings, with the exact prompts in Appendix A.1. The supplementary material contains the time-stamped specifications, the generation code, the analysis code that computes every reported table and figure from the generated outputs, every analysis output, and the generated outputs of the pre-specified crossing, its controls and the mechanism experiments; the remaining generated outputs will be released with the paper.

AI Use Statement

Generative AI assisted experimental design, implementation, debugging, analysis, literature search, figure production and manuscript editing. The authors checked the cited sources and take responsibility for the experiments, analyses and submitted text. The supplementary material provides the analysis code and its outputs.

Ethics Statement

The study evaluates public reasoning benchmarks and model outputs and contains no private participant data. Its main risk is interpreting interface-dependent measurements as intrinsic model capability; we report the evaluation choices that affect that interpretation.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Milad Aghajohari, Kamran Chitsaz, Amirhossein Kazemnejad, Sarath Chandar, Alessandro Sordoni, Aaron Courville, and Siva Reddy. The Markovian thinker: Architecture-agnostic linear scaling of reasoning. In *International Conference on Learning Representations*, 2026.

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

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

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

Francesca Carlon, Vincent Ginis, and Andres Algaba. Shorter reasoning, earlier answers? An evaluation of reasoning interfaces. arXiv preprint arXiv:2608.03401, 2026.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. arXiv preprint arXiv:2507.02856, 2025.

Haoyang Chen, Yi Liu, Jianzhi Shao, Tao Zhang, Chengfu Huo, and Wei Hu. How do answer tokens read reasoning traces? self-reading patterns in thinking LLMs for quantitative reasoning. arXiv preprint arXiv:2604.19149, 2026a.

Mingguang Chen, Bo Qu, and Licheng Wang. The calibration floor: Format repair can masquerade as self-correction at small-to-mid scale. arXiv preprint arXiv:2608.04355, 2026b.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, 2025.

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.

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 with Certaindex. In *Advances in Neural Information Processing Systems*, 2025.

Ankit Goyal and Jaideep Ray. Mind the cap: Output-budget regimes change the measured multilingual reasoning gap. arXiv preprint arXiv:2608.04160, 2026.

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.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling*, 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Junquan Huang, Haotian Wu, Yubo Gao, Yibo Yan, Junyan Zhang, Yonghua Hei, Song Dai, Jie Zhang, Puay Siew Tan, and Xuming Hu. EffiReason-Bench: A unified benchmark for evaluating and advancing efficient reasoning in large language models. arXiv preprint arXiv:2511.10201, 2025.

Hwiyeol Jo, Joosung Lee, Jaehone Lee, Sang-Woo Lee, Joonsuk Park, and Kang Min Yoo. Finding answers in thought matters: Revisiting evaluation on large language models with reasoning. arXiv preprint arXiv:2510.14773, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Yongchan Kwon, Shang Zhu, Federico Bianchi, Kaitlyn Zhou, and James Zou. ReasonIF: Large reasoning models fail to follow instructions during reasoning. arXiv preprint arXiv:2510.15211, 2025.

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.

Jiakun Li, Xingwei He, Kefan Li, Hongzheng Chai, Hongyue Yu, and Yuan Yuan. Efficient test-time scaling via temporal reasoning aggregation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13304–13318. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.651.

Junyan Li, Wenshuo Zhao, Yang Zhang, and Chuang Gan. Steering LLM thinking with budget guidance. arXiv preprint arXiv:2506.13752, 2025.

Baohao Liao, Hanze Dong, Yuhui Xu, Doyen Sahoo, Christof Monz, Junnan Li, and Caiming Xiong. Fractured chain-of-thought reasoning. In *Conference on Language Modeling*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.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.

MathArena. MathArena: Mathematical reasoning evaluation, 2026. URL `https://matharena.ai/`.

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

Dehai Min, Giovanni Vaccarino, Huiyi Chen, Yongliang Wu, Gal Yona, and Lu Cheng. Stop when reasoning converges: Semantic-preserving early exit for reasoning models. arXiv preprint arXiv:2605.17672, 2026.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

João Paulo Nogueira, Wentao Sun, Alonso Silva, and Laith Zumot. Certainty-guided reasoning in large language models: A dynamic thinking budget approach. arXiv preprint arXiv:2509.07820, 2025.

Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Optimizing anytime reasoning via budget relative policy optimization. arXiv preprint arXiv:2505.13438, 2025.

Qwen Team. Qwen3.5 model cards. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026. Released model family documentation.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. arXiv preprint arXiv:2311.12022, 2023. URL `https://arxiv.org/abs/2311.12022`.

Pavel Rumiantsev, Soumyasundar Pal, Yingxue Zhang, and Mark Coates. FEval-TTC: Fair evaluation protocol for test-time compute. arXiv preprint arXiv:2511.01203, 2025.

Ranajoy Sadhukhan, Zhuoming Chen, Haizhong Zheng, Yang Zhou, Emma Strubell, and Beidi Chen. Kinetics: Rethinking test-time scaling laws. In *Advances in Neural Information Processing Systems*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314, 2024.

Sangjun Song, Minjae Oh, Seungkyu Lee, Sungmin Jo, and Yohan Jo. ThinkBrake: Efficient reasoning via log-probability margin guided decoding. arXiv preprint arXiv:2510.00546, 2025.

Ian Su, Gaurav Purushothaman, Jey Narayan, et al. Broken chains: The cost of incomplete reasoning in LLMs. arXiv preprint arXiv:2602.14444, 2026.

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 27250–27268. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1256.

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.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236, 2024.

Hanyang Wang and Mingxuan Zhu. The detection–extraction gap: Models know the answer before they can say it. arXiv preprint arXiv:2604.06613, 2026.

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of LLM reasoning strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a.

Xi Wang, James McInerney, Lequn Wang, and Nathan Kallus. Entropy after ⟨/Think⟩ for reasoning model early exiting. arXiv preprint arXiv:2509.26522, 2025.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: Risk control for reasoning on a compute budget. arXiv preprint arXiv:2602.03814, 2026.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024b.

Tsung-Han Wu, Mihran Miroyan, David M. Chan, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. Are large reasoning models interruptible? arXiv preprint arXiv:2510.11713, 2025.

Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. Scalable chain of thoughts via elastic reasoning. arXiv preprint arXiv:2505.05315, 2025.

Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. InftyThink: Breaking the length limits of long-context reasoning in large language models. In *International Conference on Learning Representations*, 2026.

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

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

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations*, 2025.

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

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

Shlomo Zilberstein. Using anytime algorithms in intelligent systems. *AI Magazine*, 17(3):73–83, 1996.

A Rule–Interface Crossing

This appendix gives the crossing's protocol, complete results, replications and controls; Appendix B the follow-up experiments that locate the mechanism, compare the rules and test the scope; and Appendix C the pairing and reproduction checks and the audit of answer procedures in published stopping methods. Terms follow Section 2; the interfaces are defined below.

What was fixed in advance. Each result is labeled with its tier where it is reported. The 27 interactions of Section 2 form the confirmatory family. The cap cells, the seven replication cells, the MATH-500 cell and the controls form separately pre-specified families, each fixed in a time-stamped specification before its runs. Follow-up experiments (Appendix B) were specified before their own runs but after earlier results had been seen; among them are the prospective applications of the reporting checklist to Ministral-3-8B and to the MATH-500 cell. Three later checks (the token-level probe, the prefix from another problem and the second Ministral-3-8B check) finished after the time their specifications set for this version or were specified for a later one; they are reported here as late checks. Post hoc analyses, among them the stored-trial contrasts, the comparison with a smaller cap and the application of the checklist to the crossing cells, are marked as such. Every specification is time-stamped and included in the supplementary material; none was registered with a third party, so the time stamps, not an external registry, document the order.

A.1 Protocol and complete results

Base trajectories and stop points. The base trajectory of each problem and seed is drawn from a *source bank* of trajectories sampled before this study: ordinary sampling under the source allowance $B - 1$K, requested in chunks, ending either natively or at the horizon. The bank and the staged cascade come from earlier studies of the answer stage and are used here unchanged; the bank also carries a diagnostic parser arm and a cascade arm on those same trajectories, which the replication gate below uses. The reasoning region of a trajectory runs up to its first end-of-thinking token, or to its end if it has none, and a *boundary* is any position in that region that follows a token whose text contains a blank line. Boundaries found from single-token decoding agree exactly with blank lines counted in the text on every trajectory ($10{,}054$, $10{,}954$ and $4{,}344$ on AMC23, SMT25 and AIME25), and the decoded text of every one of the $1{,}954$ prefix stop points is a prefix of the stored source text. AMC23 trajectories have a median of 53 boundaries, SMT25 42 and AIME25 23.

Controllers. *Adapted EAT-v1* follows Algorithm 1 of the first arXiv version of Wang et al. (2025) (`https://arxiv.org/pdf/2509.26522v1`); the later v3 (`https://arxiv.org/pdf/2509.26522v3`) divides V_n by $1 - (1 - \alpha)^n$ and no longer lists the fixed warm-up condition. Our EAT results, in particular the matched threshold's exits at the end of warm-up, apply to this v1 rule. At boundary n the prefix is extended by the end-of-thinking token and one newline, and h_n is the entropy of the next-token distribution over the full vocabulary. With $M_0 = V_0 = 0$, $M_n = (1 - \alpha)M_{n-1} + \alpha h_n$ and $V_n = (1 - \alpha)V_{n-1} + \alpha(h_n - M_n)^2$, the rule exits at the first $n \geq 4/\alpha$ with $V_n < \delta$; $\alpha = 0.2$, so the warm-up is 20 checks and there is no bias correction. The paper sweeps δ rather than fixing it; we evaluate $\delta = 2^{-k}$ for $k \in \{1, 3, \ldots, 15\}$ and, for the inferential family only, designate per cell the threshold whose saving of reasoning tokens is closest to 15%, a target inside the 12–22% range the paper reports; the *matched* EAT (mEAT in Figure 2) is the threshold whose saving is closest to AC's. The saving is one minus the ratio of mean reasoning tokens (tokens before the first end-of-thinking token, or the whole source if there is none) under the rule and under the full trace; ties go to the smaller threshold. The rule uses stop positions only and was applied before any interface was scored. The savings of the eight thresholds from 2^{-1} to 2^{-15} are 47.3, 43.2, 32.9, 21.9, 10.5, 3.4, 0.6 and 0.1% on AMC23 16K, 31.1, 28.0, 18.4, 11.4, 5.1, 1.2, 0.3 and 0.1% on SMT25, and 25.5, 22.0, 13.0, 7.8, 3.8, 1.4, 0.7 and 0.3% on AIME25, so the rule selects 2^{-9}, 2^{-5} and 2^{-5} (2^{-5} at 8K and 2^{-9} at 32K). Loose thresholds stop at the end of the warm-up rather than through the entropy signal: at 2^{-1} and 2^{-3}, 97.5–99.0% and 57.9–76.9% of the fired trajectories exit at the first eligible check (the 20th boundary), at 2^{-5} 15.8–31.3%, and at 2^{-7} or below at most 3.1%. The thresholds at which the signal chooses most stops save at most 32.9% of reasoning tokens on AMC23 16K, 18.4% on SMT25 and 13.0% on AIME25, less than AC's 51.5, 32.7 and 25.9%: no threshold of this adapted v1 rule both matches AC's saving and selects its stops by entropy. *Answer Convergence* follows Liu & Wang (2025): at each boundary

the prefix is extended by the end-of-thinking token and the string "The final answer is \boxed{", a greedy continuation of at most 48 tokens is read with the task's canonical candidate reader, and the rule stops at the boundary that completes $k = 10$ consecutive identical candidates. The original checks at sentence boundaries; ours are coarser, so $k = 10$ spans more text here. *s1-style forcing* follows the minimum-budget mechanism of Muennighoff et al. (2025): the first two times the model emits the end-of-thinking token it is removed and "Wait" is appended, and sampling continues with source settings; afterwards the model ends on its own or reaches the source allowance. Trajectories that never try to end their reasoning are unchanged. The maximum-budget half of budget forcing, the end-of-thinking token followed by an answer prompt, is what native closure does for every policy.

EAT with the probe prefix of its experiments. Our probe follows EAT-v1's Algorithm 1 and Equation 5, the entropy directly after the end-of-thinking token and a newline; the paper's experiments add the prefix "Final answer: " for all models, and after exiting EAT appends the end-of-thinking token and "Final answer:" and generates to the end of the sequence. A follow-up reprobes every boundary of the three primary cells with the prefix and applies the unchanged rule; re-probing a sample of boundaries with the original suffix reproduces the original entropies up to batch-level numerical variation. The matched threshold stays at 2^{-1} in every cell, with savings within 0.8 points of the original and 95.0, 90.5 and 97.0% of the trajectories it cuts exiting at the first eligible check, against 97.5, 98.8 and 99.0%; it stops at the same boundary on 143 of 150, 182 of 200 and 257 of 265 trajectories, and its stored-trial accuracy is unchanged on AIME25 and AMC23 and 0.4 points lower on SMT25. The designated threshold stays at 2^{-5} on AIME25 and 2^{-9} on AMC23 and moves to 2^{-7} on SMT25. With the prefix the signal settles more often, so tight thresholds fire on more trajectories (on AMC23, 130 instead of 95 at 2^{-9}). A second follow-up reads the stops of the prefix probe's designated threshold with every interface of the controls (Table 3). The interface dependence is unchanged: closure lowers EAT's cost relative to continuation by 28.5, 15.5 and 13.3 points on AMC23, SMT25 and AIME25 against 21.5, 15.5 and 14.0 with the original probe, and the cascade costs 9.5, 13.2 and 4.0 points more than closure against 7.0, 13.6 and 5.3. Under this EAT-v1 adaptation, the prefix changes where tight thresholds stop while preserving the concentration of matched-threshold exits at warm-up and the observed interface dependence.

Table 3: **With the probe prefix of EAT's experiments, the dependence on the interface is unchanged.** EAT's designated threshold under the original probe and under the "Final answer: " probe: stop cost (pp) under continuation and native closure, and the closure-versus-continuation and cascade-versus-closure interactions, with pointwise 95% intervals.

Cell	Probe	Threshold	Fired	Cost, continuation	Cost, closure	Closure − continuation	Cascade − closure
AMC23 16K	original	2^{-9}	95	+27.0 [18.5, 36.0]	+5.5 [2.0, 9.5]	+21.5 [12.5, 31.0]	−7.0 [−13.5, −0.5]
	"Final answer: "	2^{-9}	130	+37.0 [27.5, 46.5]	+8.5 [3.5, 14.0]	+28.5 [18.5, 39.0]	−9.5 [−16.0, −3.0]
SMT25 32K	original	2^{-5}	176	+53.6 [42.6, 64.2]	+38.1 [29.4, 47.2]	+15.5 [7.9, 23.4]	−13.6 [−21.1, −6.8]
	"Final answer: "	2^{-7}	155	+45.3 [35.5, 55.5]	+29.8 [21.9, 38.1]	+15.5 [9.1, 22.3]	−13.2 [−20.4, −6.4]
AIME25 16K	original	2^{-5}	53	+22.0 [9.3, 36.0]	+8.0 [2.0, 16.0]	+14.0 [4.0, 26.7]	−5.3 [−13.3, 2.0]
	"Final answer: "	2^{-5}	68	+22.0 [9.3, 36.0]	+8.7 [2.7, 16.0]	+13.3 [3.3, 26.0]	−4.0 [−10.7, 2.0]

Interfaces. The three generating interfaces spend at most the answer allowance of $1{,}024$ tokens, injected tokens included, and every interface returns a natively terminated source unchanged. *Continuation* issues one request with source sampling and no suffix. *Native closure* appends, to a source whose thinking block is open, "Considering the limited time by the user, I have to give the solution based on the thinking directly now.", the end-of-thinking token and a blank line (25 tokens), and decodes with $T = 0$; a source whose block is closed but was cut by the horizon is continued with $T = 0$ and no suffix. For Ministral-3-8B the package is the same sentence followed by that model's end-of-thinking token, and for DeepSeek-R1-Distill-Llama-8B its end-of-thinking token alone; native closure is an applied interface, not an observed self-termination. *Strict open* and *strict closed* make one greedy call with the strict final-answer instruction, inside the open block or after the closure package (Appendix A.3). The cascade is defined below. *The stored boxed trial* is not generated separately: it is the trial answer that Answer Convergence records at every boundary (end-of-thinking token, "The final answer is \boxed{", greedy, at most 48 tokens, read by the task's candidate reader), taken at the stop point of whichever policy is being read. For a full trace that the horizon cut it is taken at the last checked blank-line boundary; for a trace that ended on its own it is the trace's own

answer. Reading instead at the exact horizon endpoint is reported separately in Appendix A.3. The diagnostic Parser returns the source.

The staged cascade. The staged cascade executes the following stages over the cumulative source and answer text. Before rescue, after rescue, and after correction it applies the task-specific zero-token deterministic readout when no terminal candidate is present. It then samples, in order: (i) rescue if no terminal candidate is present, with $T = 0.2$ and cap 768; (ii) strict answer if still unresolved, with $T = 0$ and cap 256; (iii) correction if the latest terminal marker is invalid and at least 32 tokens remain, with $T = 0$ and cap 768; and (iv) finalization if still unresolved and at least 32 tokens remain, with $T = 0$ and cap 512. Every sampled stage receives $\min(\text{stage cap}, 1024 - \text{tokens used})$. The reader selects the most recent terminal-format candidate, and the returned text contains the source plus all answer-stage output. Unused quota is not reassigned outside this ordered cascade.

Verbatim answer-stage prompts. The suffixes are verbatim as follows, each preceded by two newline characters in the existing assistant continuation. AIME rescue: "Stop now and give the final integer answer in `\boxed{N}`."; strict answer: "Output exactly one final integer answer in `\boxed{N}` and nothing else." AMC uses the same two with "signed integer" in place of "integer". SMT rescue: "Stop now and give the final answer as exactly one expression in `\boxed{...}`."; strict answer: "Output exactly one final answer in `\boxed{...}` and nothing else." GPQA rescue: "Complete the reasoning and end with exactly one final answer in the form "The correct answer is (A)", replacing A with A, B, C, or D."; strict answer: "Return only the final answer now in the form "The correct answer is (A)", replacing A with A, B, C, or D."

The two fallback stages follow the same pattern. The invalid-answer correction names the violation and asks for one corrected value—for AIME, "Your previous boxed answer was not an integer from 0 to 999. Give only one corrected integer in `\boxed{N}`.", with the AMC, SMT and GPQA variants substituting "a signed integer", "invalid" and "The final answer format is invalid" respectively. The finalizer repeats the strict instruction and adds "Do not add explanation."

Checkpoint, sampling and scoring. The shared checkpoint is `Qwen/Qwen3.5-9B`. Source sampling uses $T = 1$, top-$p = 0.95$, top-$k = 20$, and presence penalty 1.5. The base trajectories are requested in chunks, a 6,144-token first request followed by 4,096-token requests, clipped at the source allowance; the sampler's presence penalty counts tokens generated within a request, so it restarts at each chunk boundary. Native EOS ends source generation. AIME and AMC use strict task parsers and SMT the MathArena symbolic scorer (MathArena, 2026).

Generation cap and cost accounting. The generation cap B bounds sampled and injected tokens of the source and the terminal interface together: the source may use $B - 1$K, the interface at most 1K, injected closure tokens included. Entropy probes and trial answers are reported separately as auxiliary cost and not charged to B, so that a trajectory on which a rule never fires stays identical, token for token, to the full trace; charging them would shorten a horizon-terminated trace on which the rule never fires by under 0.6% of its allowance on average. Reasoning-token savings are not total-compute savings: an answer after closure may itself reason for hundreds of tokens, which the answer allowance bounds and Table 17 counts. Probes cost one token each and trial answers 3.8 to 7.5 tokens on average; per trajectory in the original five Qwen3.5-9B cells they total 20–44 tokens for designated EAT and 60–134 for Answer Convergence.

Specification and inference. The full-trace outputs of all five cells reproduce the source bank's earlier outputs exactly (Appendix C.1). The design, the family, the designation rule and the decision rule were frozen before any result; the run was interrupted once by a technical fault and resumed under a second time-stamped specification with identical procedures, and the AMC23 cap cells were added afterwards as a separate family. The stored-trial contrasts, the stratified analysis of Section 5 and the comparison with a smaller cap were added after the primary family had been read and are post hoc. Table 6 groups the 27 primary interactions, 18 cap interactions and the replication families. In the post hoc stored-trial family (five cells, simultaneous half-width 17.0), the stored trial minus continuation excludes zero in seven of the ten cell–rule pairs, minus the cascade in four and minus native closure in two. The log failure-rate scale is not estimable at 32K, where bootstrap replicates reach 100% accuracy, and is omitted there. The primary band gives every contrast the half-width of the noisiest; as a sensitivity analysis, a studentized max-t band (critical value 3.09) detects nine of

the 27 primary interactions instead of five: closure minus continuation for both rules on AMC23 and SMT25, cascade minus closure for AC on both and for EAT on SMT25, cascade minus continuation for EAT on AMC23, and one forcing-control contrast, cascade minus continuation on AMC23. None is on AIME25.

The staged cascade below the Parser on stop points. On full traces the cascade never loses a parser-correct source (0 of 167 on SMT25). On stop points it can: for Answer Convergence on SMT25 the parser scores 23 of the 194 stop points correct and the cascade submits a correct answer for only 6 of them (12 of all 194), against 59 of all 194 for native closure and 9 for continuation. The run-time reader of the cascade does not recognize every expression the official scorer accepts inside an unfinished derivation, so it generates, and what it generates replaces what the scorer would have read.

Table 4 gives, for each of the five Qwen3.5-9B crossing cells, the full-trace accuracy under each interface and the measured stop costs of the two rules; a rule's stop-point accuracy is the full-trace accuracy minus its cost. Of the full traces, 96 of 200, 175 of 265, 36 of 150, 24 of 200 and 181 of 200 end on their own in AMC23 16K, SMT25 32K, AIME25 16K, AMC23 8K and AMC23 32K; the rest are horizon-cut.

Table 4: Among the three generating interfaces, native closure gives both rules their lowest stop cost in every cell. Full-trace accuracy (%) under each interface, and each rule's measured stop cost $C(P, F)$ against it in points, with pointwise 95% problem-cluster intervals. $[0.0, 0.0]$ marks identical outcomes on every trajectory: at EAT's late stops the stored trial returns the full trace's answer.

Cell		Continuation	Cascade	Closure	Stored trial
AMC23 16K	Full accuracy (%)	59.5	80.5	72.5	87.5
	EAT cost	$+27.0$ $[18.5, 36.0]$	$+12.5$ $[7.0, 18.5]$	$+5.5$ $[2.0, 9.5]$	$+0.0$ $[0.0, 0.0]$
	Answer Convergence cost	$+51.5$ $[39.5, 63.5]$	$+35.5$ $[25.5, 45.5]$	$+7.0$ $[1.5, 13.0]$	$+6.0$ $[1.0, 12.0]$
SMT25 32K	Full accuracy (%)	65.7	67.2	68.7	73.6
	EAT cost	$+53.6$ $[42.6, 64.2]$	$+51.7$ $[41.1, 62.3]$	$+38.1$ $[29.4, 47.2]$	$+4.2$ $[1.1, 7.9]$
	Answer Convergence cost	$+55.5$ $[44.9, 66.0]$	$+54.7$ $[43.8, 65.3]$	$+37.7$ $[29.1, 46.8]$	$+10.6$ $[6.0, 15.5]$
AIME25 16K	Full accuracy (%)	25.3	36.7	28.0	55.3
	EAT cost	$+22.0$ $[9.3, 36.0]$	$+13.3$ $[2.7, 25.3]$	$+8.0$ $[2.0, 16.0]$	$+0.7$ $[0.0, 2.0]$
	Answer Convergence cost	$+22.0$ $[9.3, 36.0]$	$+15.3$ $[6.0, 26.0]$	$+8.0$ $[2.0, 16.7]$	$+2.7$ $[0.0, 6.0]$
AMC23 8K	Full accuracy (%)	20.5	54.5	57.0	82.5
	EAT cost	$+12.5$ $[6.5, 19.5]$	$+3.5$ $[-1.0, 8.0]$	-3.5 $[-8.0, 0.5]$	$+0.0$ $[0.0, 0.0]$
	Answer Convergence cost	$+14.0$ $[6.0, 23.5]$	$+9.5$ $[3.0, 16.0]$	-3.5 $[-8.0, 0.5]$	$+3.5$ $[0.0, 8.5]$
AMC23 32K	Full accuracy (%)	93.0	98.5	95.5	98.5
	EAT cost	$+44.5$ $[35.5, 53.5]$	$+16.0$ $[10.5, 22.0]$	$+15.5$ $[8.5, 23.5]$	$+2.0$ $[0.0, 6.0]$
	Answer Convergence cost	$+79.5$ $[71.5, 87.0]$	$+48.5$ $[40.5, 56.5]$	$+27.0$ $[17.0, 38.0]$	$+10.0$ $[3.5, 18.0]$

What the stop point contains. On the prefixes where AC fires, the stored trial answer is correct more often than a fresh native closure in all ten Qwen crossing cells, the one-shot cell and the DeepSeek-R1-Distill-Llama-8B cell, and less often in both Ministral-3-8B cells; the cascade loses more ground to closure at AC's stops than at horizon cuts in most cells. Table 5 restricts the three generating interfaces to trajectories that would otherwise have ended natively; their full traces are returned unchanged by every interface, so these columns compare interfaces on the same interrupted but otherwise finished sources. Native closure scores highest in eleven of the twelve cells; on AIME25 32K the cascade is 1.4 points higher.

Table 5: **On traces that would have ended on their own, native closure answers AC's stops far better than the open interfaces.** Accuracy (%) of the three generating interfaces at AC's stops on trajectories that would otherwise have ended natively (n cut trajectories).

Cell	n	Continuation	Cascade	Closure
Qwen3.5-9B				
AMC23 16K	92	12.0	64.1	95.7
SMT25 32K	159	5.7	6.9	34.6
AIME25 16K	31	0.0	48.4	64.5
AMC23 8K	24	20.8	87.5	100.0
AMC23 32K	170	9.4	48.8	70.0
AIME25 32K	74	0.0	36.5	35.1
BRUMO25 32K	91	6.6	7.7	44.0
HMMT-Feb25 32K	68	2.9	2.9	22.1
CMIMC25 32K	66	6.1	7.6	19.7
Qwen3.5-27B				
AMC23 16K	121	9.1	61.2	87.6
Ministral-3-8B				
AMC23 16K	106	10.4	55.7	72.6
AIME25 16K	39	2.6	30.8	79.5

Complete interaction results. Figures 2 and 6 summarize the crossing. Table 6 reports every rule–interface interaction and its interval, grouped by inferential family, so that any estimate behind those figures and Sections 3–6 can be checked.

Table 6: **Complete rule–interface interaction results underlying Figures 2 and 6.** Each rule's interaction against the full trace, in points, with the interval named in the family heading (†: excludes zero). Columns compare closure with continuation, cascade with continuation, and cascade with closure; on SMT25 part of the cascade's loss at AC's stops comes from its run-time reader, which misses some expressions the official scorer accepts (Appendix A.1).

Cell	Rule	Closure−Cont.	Cascade−Cont.	Cascade−Closure
Qwen3.5-9B, primary: 27 interactions, half-width 16.0 pp				
AMC23 16K	EAT	$+21.5$ [5.5, 37.5]†	$+14.5$ [-1.5, 30.5]	-7.0 [-23.0, 9.0]
	AC	$+44.5$ [28.5, 60.5]†	$+16.0$ [0.0, 32.0]	-28.5 [-44.5, -12.5]†
	s1	$+5.0$ [-11.0, 21.0]	$+8.0$ [-8.0, 24.0]	$+3.0$ [-13.0, 19.0]
SMT25 32K	EAT	$+15.5$ [-0.5, 31.5]	$+1.9$ [-14.1, 17.9]	-13.6 [-29.6, 2.4]
	AC	$+17.7$ [1.7, 33.7]†	$+0.8$ [-15.2, 16.8]	-17.0 [-33.0, -1.0]†
	s1	-0.4 [-16.4, 15.6]	$+2.6$ [-13.4, 18.6]	$+3.0$ [-13.0, 19.0]
AIME25 16K	EAT	$+14.0$ [-2.0, 30.0]	$+8.7$ [-7.3, 24.7]	-5.3 [-21.3, 10.7]
	AC	$+14.0$ [-2.0, 30.0]	$+6.7$ [-9.3, 22.7]	-7.3 [-23.3, 8.7]
	s1	$+1.3$ [-14.7, 17.3]	$+8.0$ [-8.0, 24.0]	$+6.7$ [-9.3, 22.7]
Qwen3.5-9B, cap: 18 interactions, half-width 13.0 pp				
AMC23 8K	EAT	$+16.0$ [3.0, 29.0]†	$+9.0$ [-4.0, 22.0]	-7.0 [-20.0, 6.0]
	AC	$+17.5$ [4.5, 30.5]†	$+4.5$ [-8.5, 17.5]	-13.0 [-26.0, 0.0]
	s1	$+5.0$ [-8.0, 18.0]	$+5.5$ [-7.5, 18.5]	$+0.5$ [-12.5, 13.5]
AMC23 32K	EAT	$+29.0$ [16.0, 42.0]†	$+28.5$ [15.5, 41.5]†	-0.5 [-13.5, 12.5]
	AC	$+52.5$ [39.5, 65.5]†	$+31.0$ [18.0, 44.0]†	-21.5 [-34.5, -8.5]†
	s1	$+1.0$ [-12.0, 14.0]	$+2.0$ [-11.0, 15.0]	$+1.0$ [-12.0, 14.0]
Qwen3.5-9B, AIME25 32K: 9 interactions, half-width 13.3 pp				
AIME25 32K	EAT	$+18.0$ [4.7, 31.3]†	$+24.0$ [10.7, 37.3]†	$+6.0$ [-7.3, 19.3]
	AC	$+16.7$ [3.3, 30.0]†	$+16.0$ [2.7, 29.3]†	-0.7 [-14.0, 12.7]
	s1	$+0.0$ [-13.3, 13.3]	$+5.3$ [-8.0, 18.7]	$+5.3$ [-8.0, 18.7]
Qwen3.5-9B, additional tasks: 27 interactions, half-width 15.3 pp				
BRUMO25 32K	EAT	$+28.7$ [13.3, 44.0]†	$+2.7$ [-12.7, 18.0]	-26.0 [-41.3, -10.7]†
	AC	$+24.7$ [9.3, 40.0]†	$+0.0$ [-15.3, 15.3]	-24.7 [-40.0, -9.3]†
	s1	-0.7 [-16.0, 14.7]	$+0.7$ [-14.7, 16.0]	$+1.3$ [-14.0, 16.7]
HMMT-Feb25 32K	EAT	$+10.7$ [-4.7, 26.0]	$+0.7$ [-14.7, 16.0]	-10.0 [-25.3, 5.3]
	AC	$+8.0$ [-7.3, 23.3]	-4.0 [-19.3, 11.3]	-12.0 [-27.3, 3.3]
	s1	$+1.3$ [-14.0, 16.7]	$+4.7$ [-10.7, 20.0]	$+3.3$ [-12.0, 18.7]
CMIMC25 32K	EAT	$+3.0$ [-12.3, 18.3]	-0.5 [-15.8, 14.8]	-3.5 [-18.8, 11.8]
	AC	$+3.0$ [-12.3, 18.3]	-1.5 [-16.8, 13.8]	-4.5 [-19.8, 10.8]
	s1	$+0.5$ [-14.8, 15.8]	$+0.0$ [-15.3, 15.3]	-0.5 [-15.8, 14.8]
Qwen3.5-27B: 9 interactions, half-width 15.5 pp				
AMC23 16K	EAT	$+34.0$ [18.5, 49.5]†	$+27.0$ [11.5, 42.5]†	-7.0 [-22.5, 8.5]
	AC	$+48.0$ [32.5, 63.5]†	$+29.5$ [14.0, 45.0]†	-18.5 [-34.0, -3.0]†
	s1	$+6.0$ [-9.5, 21.5]	$+10.5$ [-5.0, 26.0]	$+4.5$ [-11.0, 20.0]
Ministral-3-8B: 18 interactions, half-width 17.5 pp				

Cell	Rule	Closure−Cont.	Cascade−Cont.	Cascade−Closure
AMC23 16K	EAT	+13.5 [−4.0, 31.0]	+10.5 [−7.0, 28.0]	−3.0 [−20.5, 14.5]
	AC	+25.0 [7.5, 42.5]†	+24.0 [6.5, 41.5]†	−1.0 [−18.5, 16.5]
	s1	+2.5 [−15.0, 20.0]	+2.5 [−15.0, 20.0]	+0.0 [−17.5, 17.5]
AIME25 16K	EAT	+12.7 [−4.8, 30.2]	+4.0 [−13.5, 21.5]	−8.7 [−26.2, 8.8]
	AC	+13.3 [−4.2, 30.8]	+4.7 [−12.8, 22.2]	−8.7 [−26.2, 8.8]
	s1	+0.0 [−17.5, 17.5]	+0.0 [−17.5, 17.5]	+0.0 [−17.5, 17.5]
Qwen3.5-9B, one-shot sources: family of six, half-width 15.5 pp (cascade−cont. and s1 pointwise)				
AMC23 16K	EAT	+21.5 [6.0, 37.0]†	+16.5 [9.5, 24.0]	−5.0 [−20.5, 10.5]
	AC	+38.0 [22.5, 53.5]†	+27.0 [14.5, 39.5]	−11.0 [−26.5, 4.5]
	s1	+3.0 [0.5, 5.5]	+7.0 [3.5, 10.5]	+4.0 [1.5, 7.0]
Qwen3.5-9B, GPQA-Diamond 16K: 11 contrasts, half-width 7.4 pp				
GPQA-Diamond	EAT	+19.9 [12.5, 27.3]†	+1.9 [−5.5, 9.3]	−18.0 [−25.4, −10.6]†
	AC	+24.3 [16.9, 31.7]†	+2.5 [−4.9, 9.9]	−21.8 [−29.2, −14.4]†
	s1	+0.7 [−6.7, 8.1]	+0.9 [−6.5, 8.3]	+0.2 [−7.2, 7.6]
DeepSeek-R1-Distill-Llama-8B, one-shot sources: pointwise 95% intervals				
AMC23 16K	EAT	+0.0 [0.0, 0.0]	+0.5 [0.0, 1.5]	+0.5 [0.0, 1.5]
	AC	+8.5 [3.5, 13.5]	+14.0 [8.0, 20.5]	+5.5 [−0.5, 11.5]

The primary interaction is also present on the log failure-rate scale for Answer Convergence on AMC23: closure versus continuation is -0.59 $[-0.94, -0.33]$ and cascade versus closure is $+0.81$ $[+0.50, +1.22]$ (pointwise). At 32K this scale has bootstrap replicates at 100% accuracy and is not estimated.

A.2 Replications

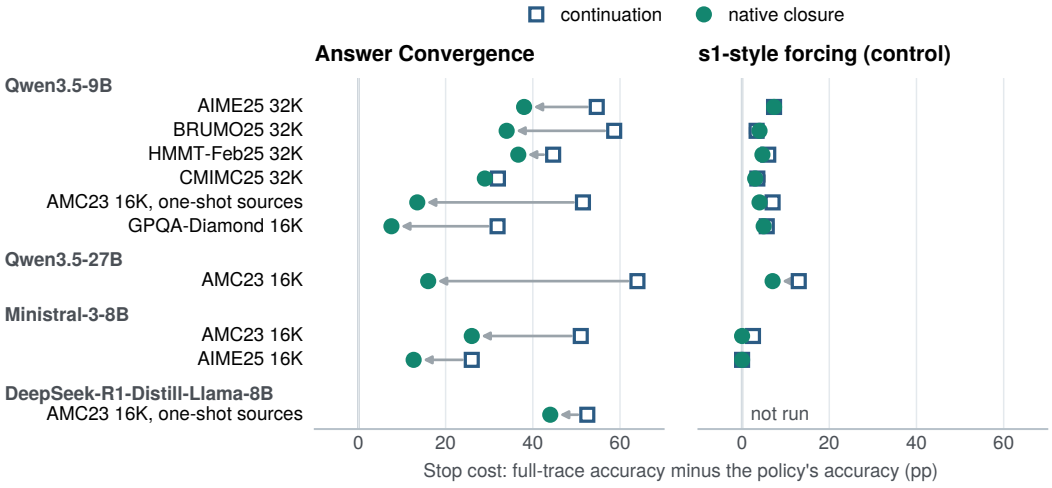

Figure 6: **Native closure lowers AC's stop cost relative to continuation in every additional cell, while the s1-style forcing control changes little.** Stop cost, full-trace accuracy minus the policy's accuracy, under continuation (squares) and native closure (circles); arrows run from continuation to closure. For DeepSeek-R1-Distill-Llama-8B only the source endpoint, the designated EAT threshold and AC were run. The primary cells and AMC23 32K are in Figure 2; AMC23 8K is in Tables 4 and 6.

Seven additional cells repeat the rule–interface crossing with five seeds per problem: AIME25 at 32K, which reuses the existing AIME25 32K trajectories; and BRUMO25, HMMT-Feb25 and CMIMC25 at 32K, AMC23 on Qwen3.5-27B, and AMC23 and AIME25 on Ministral-3-8B, which sample new base trajectories with the same controller. Figure 6 shows their stop costs under continuation and native closure; separately labeled interaction families appear in Table 6. The family sizes are nine for AIME25 32K and Qwen3.5-27B, 27 for the three additional tasks jointly, and 18 for the two Ministral cells; their simultaneous half-widths are 13.3, 15.5, 15.3 and 17.5 points, respectively. Qwen3.5-27B designates $\delta = 2^{-9}$ by the same outcome-blind rule, saving 18.4% of reasoning tokens. Their stored-trial and source-state comparisons are summarized in Appendix A.1.

A larger cell specified before its generation. After every other family had been analyzed, one more cell was specified before any of its generation, with the design, controllers, designation rule and family of the held-out cells unchanged and a directional prediction fixed in advance: MATH-500 (Lightman et al., 2023) level 5 at 32K, the 127 level-5 problems whose official answer the scorer reads back (seven are excluded by that rule), five seeds and 635 trajectories. MATH-500 was used in the early development of the cascade's symbolic prompts, which if anything favors the cascade. Here 92.4% of the full traces end on their own, and the three generating interfaces are within 1.4 points of each other on them (85.2 to 86.6%). AC fires on 570 trajectories and saves 48.6% of the reasoning tokens; its stops cost 67.4 points under the staged cascade and 27.2 after native closure (Table 7), so the cascade costs 40.2 points more ($[33.4, 46.8]$), as predicted. With 127 problems the cell's nine-interaction simultaneous band has a half-width of 7.9 points; the four closure contrasts of AC and the designated EAT threshold exclude zero, and none of the forcing-control contrasts does.

Table 7: **On a 127-problem cell fixed before generation, where most full traces end on their own, the stop still costs far more under the cascade and continuation than after native closure.** MATH-500 level 5, Qwen3.5-9B, 32K: firing, reasoning-token saving (%), stop cost (pp) under each generating interface and the cascade's extra cost over native closure, with pointwise 95% problem-cluster intervals.

Rule	Fired	Saving	Continuation	Cascade	Closure	Cascade − closure
AC	570	48.6	+69.9 [63.6, 75.9]	+67.4 [61.6, 72.9]	+27.2 [21.6, 33.2]	+40.2 [33.4, 46.8]
EAT, designated (2^{-9})	339	11.2	+36.2 [30.6, 41.9]	+32.3 [27.1, 37.5]	+15.4 [11.0, 20.2]	+16.9 [11.7, 22.2]
s1-style forcing	595	–	+2.8 [1.1, 4.9]	+2.2 [0.8, 3.9]	+2.0 [0.8, 3.8]	+0.2 [−0.8, 1.3]

A.3 Controls

Protocol of the controls. The controls were specified after the primary and cap families had been read and before any of their own results; they add arms and stop points to the existing trajectories of three cells (AMC23 16K, SMT25 32K, AMC23 32K) without resampling the sources. *Closure × cascade* adds three arms to the stop points of Full trace, the designated EAT threshold and Answer Convergence: *strict open*, a single greedy request with the strict final-answer instruction (the cascade's finalizer) appended inside the open thinking block, skipped when the stop point already carries a terminal candidate; *strict closed*, the same request after the closure package has been injected; and *closure then cascade*, the full cascade after that injection. Together with the crossing's cascade (open) they form a 2×2 in which closing the block and using several stages vary separately; on a stop point whose block the model had already closed, the closed arms coincide with the open ones by construction. The 25 closure tokens are injected on top of the 1K answer allowance in the closed arms. *Depth-matched random cuts* take the trajectories a rule cuts and the relative depths at which it cuts them (stop position over reasoning length), permute those depths across the trajectories within fixed blocks of trajectories, and cut each trajectory at the boundary nearest its reassigned depth; the rule and its random counterpart therefore cut the same trajectories at the same distribution of relative depths, and differ only in *which* trajectory is cut *where*. We draw five seeded shuffles, the first specified with the controls and four more in a later specification written before they were run, and report their mean; the bootstrap resamples problems jointly over the shuffles. *The empty trace* hands every interface the question with no reasoning at all. *The allowance sweep* repeats continuation and native closure with 2K and 4K answer allowances on every rule-cut, random-cut and empty stop point that leaves room for 4K under the cap. *Matched token budgets*, a separate run on the three primary cells, cut every trajectory at T reasoning tokens, with T chosen per cell and rule from token counts alone so that the mean number of reasoning tokens equals the rule's ($T = 6{,}124$ for Answer Convergence and $12{,}855$ for EAT on AMC23; $17{,}045$ and $21{,}901$ on SMT25; $10{,}804$ and $12{,}948$ on AIME25); because a budget does not fall on a stored boundary, a boxed trial is generated at those cuts, and also at the true end of every horizon-cut full trace. All contrasts use a problem-cluster bootstrap with 20,000 draws and are reported pointwise.

Closing the block versus adding stages. Whether stages help depends on the block state (Table 8). On the two AMC23 cells the cascade helps while the block is open and hurts after closure, at every rule stop and at the full traces; on SMT25 it helps in both states at the rule stops. Over all trajectories, closing the block raises AC-stop accuracy by 49.0, 52.3 and 46.3 points averaged over the two stage

structures (pointwise intervals exclude zero), and the closure-by-staging interaction at AC's stops is -41.0 and -42.5 points on the AMC23 cells against $+7.9$ on SMT25. With an empty trace closing the block does not matter (closure -4.5 and $+0.2$ points, intervals including zero), while staging still adds 25.0 points on AMC23 ($[14.5, 36.3]$).

Table 8: **Closing the block raises accuracy at every rule stop; staging helps only while the block is open on AMC23.** The four accuracy columns (%) are one strict call and the staged cascade with the block open or closed, restricted to the n sources whose block is open. The closure column is its marginal effect over all trajectories; the interaction is (closed cascade − closed call) − (open cascade − open call); both in points with pointwise 95% intervals.

			Open		Closed		Contrasts [CI]	
Cell	Policy	n	Call	Cascade	Call	Cascade	Closure	Interaction
AMC23 16K	Full	87	29.9	56.3	85.1	67.8	+14.5 [8.8, 20.5]	−19.0 [−30.5, −8.0]
	EAT	158	28.5	60.1	90.5	77.2	+31.3 [25.8, 37.0]	−35.5 [−45.5, −25.5]
	AC	195	12.8	43.6	84.1	72.8	+49.0 [41.5, 56.5]	−41.0 [−52.0, −30.0]
	Empty	200	20.0	45.5	16.0	40.5	−4.5 [−11.0, 1.8]	−1.0 [−11.5, 10.0]
SMT25 32K	Full	85	9.4	8.2	41.2	43.5	+10.8 [6.0, 16.2]	+1.1 [−1.1, 3.4]
	EAT	244	6.6	9.0	57.8	75.0	+54.0 [46.0, 61.7]	+13.6 [6.4, 21.1]
	AC	248	4.8	6.9	56.5	66.9	+52.3 [43.8, 60.6]	+7.9 [0.8, 15.1]
	Empty	265	6.0	6.4	3.8	9.1	+0.2 [−4.3, 4.9]	+4.9 [−3.8, 14.7]
AMC23 32K	Full	17	35.3	82.4	88.2	76.5	+2.0 [0.0, 4.8]	−5.0 [−11.5, 0.0]
	EAT	130	26.2	73.1	96.2	83.9	+26.3 [21.3, 31.5]	−38.5 [−47.0, −30.5]
	AC	189	15.9	47.1	87.3	73.5	+46.3 [39.0, 53.5]	−42.5 [−51.5, −33.5]
	Empty	200	20.0	45.5	16.0	40.5	−4.5 [−11.0, 1.8]	−1.0 [−11.5, 10.0]

More room after stopping. The allowance sweep of the controls reads the rule, depth-matched random (first shuffle) and empty-trace stops that leave room for a 4K answer under the cap, with the same trajectory sets at all three allowances; it agrees with the source-fixed experiment of Appendix B.3, which Figure 4 shows.

Two hundred shuffles under the stored trial. Under the stored trial a shuffle costs no generation, because a trial answer is stored at every boundary, so the depth-matched comparison can be repeated. Over 200 shuffles, on the trajectories the rule cuts, the trial answer at Answer Convergence's own stop points is correct 88.8% of the time on AMC23 16K against a shuffle mean of 81.4%, 71.1% against 63.7% on SMT25, 73.9% against 60.2% on AIME25 and 88.5% against 80.4% on AMC23 32K, above every one of the 200 shuffles in each cell. EAT's stop points are above every shuffle on SMT25 (85.8% against 80.6%), above 94.5% of them on AIME25 (96.2% against 92.0%), and indistinguishable from a shuffle on AMC23 (97.9% against 97.9% at 16K, 96.5% against 96.4% at 32K). Answer Convergence, which stops where trial answers agree, therefore does pick better places than its depth alone explains, by 7–14 points on the trajectories it cuts; that advantage is visible under closed readouts and absent under open ones (Table 9).

Location gains under closed readouts. The five generated shuffles show the same split. Averaged over them, AC's own stops beat depth-matched random cuts by 3.8 to 8.8 points under each of the four closed readouts (native closure, strict closed, closure then cascade and the stored trial) in all three control cells; the difference is positive in every shuffle for all 12 cell–readout pairs, and its pointwise interval excludes zero in 11 of them. Under the three open readouts (continuation, the cascade and strict open) it lies between -3.1 and $+0.7$ points, and no interval excludes zero. EAT's location adds nothing on either AMC23 cell; on SMT25 it helps under two closed readouts, closure then cascade ($+5.5$, $[3.0, 8.1]$) and the stored trial ($+3.5$, $[1.2, 6.0]$).

Rule stop points versus matched fixed-token cuts. For each rule, T is chosen without answer outcomes so the fixed cut has nearly the same mean reasoning-token length as the rule. It still cuts a different set of trajectories, shown in separate count columns in Table 9. Differences are rule minus fixed cut on the same problems and seeds. On AMC23, Answer Convergence is behind the fixed cut under continuation and the cascade but ahead under native closure; on SMT25, EAT is behind under

all three generating interfaces. Length matching therefore retains interface-dependent comparisons without making a fixed cut equivalent to a rule-selected stop. The stored trial uses the last stored blank-line boundary of every horizon-cut full trace, matching the crossing's primary definition. As a sensitivity readout, taking the stored trial at the exact end of the full trace instead raises its full-trace accuracy from 87.5 to 91.0% on AMC23, 73.6 to 77.4% on SMT25, and 55.3 to 58.0% on AIME25.

Table 9: AC's location gains are detected under closed readouts; against a length-matched fixed cut the sign depends on the interface. Rule minus a comparison cut, in accuracy points with pointwise 95% problem-cluster intervals. Upper: mean over five seeded shuffles, each reallocating the rule's relative stop depths among the trajectories it cuts. Lower: a fixed token cut matches the rule's mean reasoning length. Rule and comparison counts are separate columns. The controls match different aspects of length and are not pooled. Under the stored trial the rule minus the fixed cut is -1.0, $+3.8$ and $+2.0$ points for AC and -3.0, $+3.8$ and -1.3 for EAT on AMC23, SMT25 and AIME25, every pointwise interval including zero.

Cell	Rule	Rule n	Comparison n	Continuation	Cascade	Closure
Relative-depth shuffle on the rule-cut trajectories, mean of five shuffles						
AMC23 16K	AC	169	169	−0.8 [−3.7, 1.9]	−3.1 [−7.8, 2.4]	+6.7 [3.3, 10.5]
	EAT	95	95	−0.3 [−4.2, 3.6]	−1.2 [−5.3, 2.7]	+0.4 [−1.8, 2.7]
SMT25 32K	AC	194	194	−1.7 [−4.3, 0.8]	−1.4 [−3.9, 1.4]	+3.8 [0.1, 7.9]
	EAT	176	176	−2.4 [−5.9, 0.9]	−3.0 [−6.4, 0.2]	−2.8 [−6.3, 0.7]
AMC23 32K	AC	183	183	+0.7 [−2.9, 4.5]	−1.4 [−6.7, 4.1]	+6.9 [2.6, 11.4]
	EAT	114	114	−0.5 [−3.6, 2.5]	+2.5 [−3.1, 7.7]	−1.2 [−4.2, 1.6]
Fixed-token cut at the rule's mean reasoning length						
AMC23 16K	AC	169	177	−11.5 [−21.0, −2.5]	−9.5 [−18.0, −1.0]	+8.0 [3.0, 13.5]
	EAT	95	116	−17.0 [−26.0, −8.5]	−6.5 [−15.5, 2.5]	+1.5 [−2.5, 6.0]
SMT25 32K	AC	194	211	−13.2 [−23.8, −3.4]	−16.2 [−27.2, −5.7]	−3.0 [−12.1, 5.7]
	EAT	176	163	−28.3 [−40.8, −16.2]	−28.7 [−41.1, −16.6]	−14.7 [−25.3, −4.9]
AIME25 16K	AC	88	136	−6.7 [−16.0, 0.7]	−2.7 [−11.3, 6.7]	+3.3 [−0.7, 8.0]
	EAT	53	121	−17.3 [−30.0, −6.0]	−6.7 [−16.7, 2.7]	−2.7 [−10.0, 3.3]

Answer Convergence at sentence ends. Sentence-level checks move Answer Convergence to much earlier prefixes. A follow-up repeats the rule on AMC23 16K with a check after every sentence end as well as every blank line, about 14 times as many positions; trial answers are the stored ones at blank lines and come from the same trial request elsewhere. With $k = 10$ the rule fires on all 200 trajectories at a median of 3.3% of the reasoning, saving 95.5% of reasoning tokens against 51.5% at blank lines, and it stops earlier than the blank-line rule on all 169 trajectories where both fire. Ten agreeing trials at that depth mostly repeat an early guess: the stored trial at these stops is correct 34.0% of the time, against 88.8% at the blank-line rule's stops. The stop costs 30.0 points under native closure, which can still reason within its 1K allowance, and 51.5 to 59.5 under the other interfaces and the stored trial. The cost still depends on the interface: closure versus continuation gives $+27.5$ points ($[10.5, 44.5]$) and the cascade versus closure -21.5 ($[-32.0, -10.5]$), against $+44.5$ and -28.5 at blank lines. The strict-pair interaction is not detected: every stop is eligible, and strict closed beats strict open on them by 21.5 points ($[10.0, 33.5]$), no more than the 24.0 points it adds over all full traces, so $I = -2.5$ ($[-18.5, 13.0]$) against $+45.5$ at blank lines. The blank-line adaptation thus saves fewer reasoning tokens than the original granularity but shows the larger closure effect, and the larger effect occurs at prefixes whose short trial is more often correct.

B Follow-Up Mechanism Experiments

Each experiment in this section was specified in its own time-stamped document before its run but after earlier results had been seen, and has its own comparison family; the documents are in the supplementary material and state what had been seen. Inference uses the problem-cluster bootstrap of the crossing (five seeds kept with their problem, 20,000 draws), resampling the problems of a benchmark jointly across its cells, with max-absolute-deviation simultaneous bands, and a cell enters an analysis only if every problem, seed and arm completed.

B.1 Validity and composition

Eligibility and the three-part decomposition. A source is eligible when its thinking block is open and the common entry reader, applied once to the original text, finds no answer. On an ineligible source each arm of the strict pair either makes the same call or none, so the strict-pair contrast vanishes there; the generating interfaces need not coincide on ineligible sources, and the identity below is used for the strict pair only. For policy P let q_P be the eligible share and τ_P an interface contrast on eligible sources. The operational interaction is $I = q_{\text{AC}}\tau_{\text{AC}} - q_{\text{Full}}\tau_{\text{Full}}$; τ_{AC} is the mechanism estimand, and $\tau_{\text{AC}} - \tau_{\text{Full}}$, reported only for cells with at least 30 eligible source endpoints, describes whether early stops respond differently from deep horizon cuts. A submission is valid when the unchanged final reader extracts a legal answer; the accuracy change of an interface contrast splits symmetrically into a validity term and a conditional-correctness term, $v_1c_1 - v_0c_0 = \frac{c_1+c_0}{2}(v_1 - v_0) + \frac{v_1+v_0}{2}(c_1 - c_0)$.

Validity and composition. Table 10 splits the AC-minus-endpoint interaction of two interface contrasts into its validity term and the remainder in every crossing cell. The symbolic tasks are read with a legal-box validity rule, because their official reader falls back to the last integer of the text and would count almost every submission as valid. Closure versus continuation is carried by validity in all 12 cells; the staged cascade loses to closure mainly through validity in most cells, mainly through committing to wrong answers on AMC23 32K and Ministral-3-8B AIME25, and with the two terms nearly cancelling on AIME25 32K. For strict closed minus strict open, Figure 7 separates the composition of eligible sources from the conditional response; composition carries the larger share in nine of the ten cells where both are estimable, the exception being AIME25 16K, whose source endpoints are mostly open and unanswered.

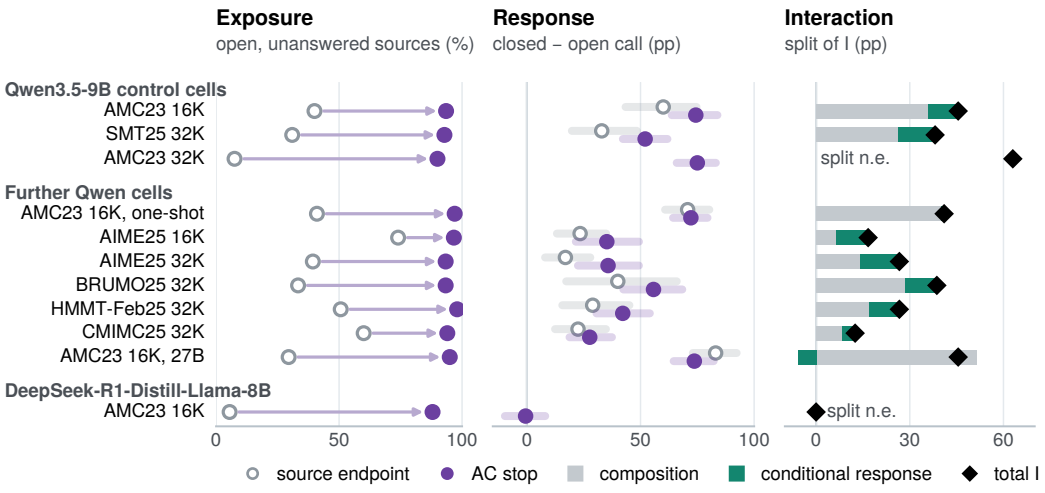

Figure 7: **Early stops mostly raise exposure: nearly every AC stop is open and unanswered, and composition carries most of the strict pair's interaction.** Strict closed minus strict open at 1K. Left: the share of sources whose block is open and which hold no answer, at the source endpoint and at AC's stops. Middle: the effect τ on those sources, with pointwise 95% intervals as shaded bars. Right: the operational interaction I split into composition $\bar{\tau}\,\Delta q$ and conditional response $\bar{q}\,\Delta\tau$; n.e.: fewer than 30 eligible endpoints.

Paired answer states. The validity decomposition is descriptive, because the two interfaces of a contrast submit on different stops. Pairing the submissions stop by stop shows what the validity term contains (Table 11). Of the 546 AC stops of the control cells, 368 move from invalid under strict open to correct under strict closed, and at 356 of them the 48-token stored trial at the same stop is also correct; the prefix text itself already scores correct on 19. A short readout of the prefix therefore yields the answer that closing the block recovers, although the closed call itself often writes a derivation first: its median length on these stops is 528 tokens, and 40% answer within 32. From continuation to native closure 276 stops move from invalid to correct, 261 of them with a correct

Table 10: Validity decomposition of the AC-minus-endpoint interaction (pp) in every crossing cell and the DeepSeek-R1-Distill-Llama-8B cell, extending Table 2 of the main text. For each contrast: total, validity term with a pointwise 95% interval, and the remainder (conditional correctness; on the symbolic tasks marked $*$ it also contains a small fallback-reader term). Unmarked rows are Qwen3.5-9B; DeepSeek is DeepSeek-R1-Distill-Llama-8B on one-shot sources. Where accuracy and validity are low (Ministral-3-8B AIME25) the percentile intervals of the two nonlinear terms are wide and asymmetric. The second contrast is cascade minus closure, the opposite sign of Table 2.

	Native closure − continuation			Native closure − staged cascade		
Cell	Total	Validity	Conditional	Total	Validity	Conditional
AMC23 16K	+44.5	+44.5 [31.0, 57.5]	+0.0	+28.5	+17.0 [11.5, 22.5]	+11.5
SMT25 32K*	+17.7	+16.9 [9.7, 24.7]	+0.8	+17.0	+16.6 [9.5, 24.4]	+0.4
AIME25 16K	+14.0	+14.0 [3.3, 26.0]	+0.0	+7.3	+5.7 [-0.4, 12.4]	+1.6
AMC23 8K	+17.5	+15.1 [5.2, 25.0]	+2.4	+13.0	+11.4 [5.2, 18.2]	+1.6
AMC23 32K	+52.5	+52.8 [41.6, 63.8]	−0.3	+21.5	+4.0 [-5.4, 13.2]	+17.5
AIME25 32K	+16.7	+16.4 [5.6, 28.8]	+0.2	+0.7	−2.4 [-11.9, 7.4]	+3.1
BRUMO25 32K*	+24.7	+21.1 [9.4, 34.8]	+3.5	+24.7	+16.8 [5.9, 29.1]	+7.9
HMMT-Feb25 32K*	+8.0	+9.3 [2.9, 16.9]	−1.3	+12.0	+12.6 [6.2, 19.6]	−0.6
CMIMC25 32K*	+3.0	+2.2 [-3.4, 8.4]	+0.8	+4.5	+4.7 [-0.6, 10.9]	−0.2
27B AMC23 16K	+48.0	+48.0 [33.5, 62.0]	+0.0	+18.5	+9.8 [1.1, 19.1]	+8.7
Ministral AMC23 16K	+25.0	+30.7 [15.7, 45.9]	−5.7	+1.0	+0.9 [-7.2, 8.7]	+0.1
Ministral AIME25 16K	+13.3	+13.7 [-20.8, 27.4]	−0.4	+8.7	+0.8 [-6.0, 9.0]	+7.9
DeepSeek AMC23	+8.5	+9.3 [4.7, 14.2]	−0.8	−5.5	−10.8 [-18.0, -3.1]	+5.3

stored trial, and native closure writes a median of 759 tokens. Almost no stop moves the other way (one and seven).

Table 11: **Closing the block turns invalid submissions into correct ones, and almost never the reverse.** Paired answer states at AC's stops (fired trajectories) of AMC23 16K, SMT25 32K and AMC23 32K, pooled; rows: state under the first interface, columns: state under the second. Valid-wrong: a legal final answer that is wrong (on SMT25 a legal boxed expression).

	Strict open → strict closed			Continuation → native closure		
From	correct	valid-wrong	invalid	correct	valid-wrong	invalid
correct	57	2	1	29	0	7
valid-wrong	0	0	0	0	0	0
invalid	368	48	70	276	4	230

A prefix from another problem. A late check replaces the retained prefix at each of AC's fired stops in the control cells by the AC stop prefix of another problem (the next fired stop, in a fixed order, of a different problem), keeping the problem statement, and reads it with the crossing's interfaces (Table 12). The closure readouts collapse: the strict closed call, correct on 60.8 to 87.4% of the stops with their own prefix, is correct on 0.0 to 7.1% with another problem's, below its 5.2 to 17.8% on the empty trace, and on 56.7 to 76.5% of the stops it submits the other problem's answer; native closure behaves alike. The readouts at AC's stops therefore depend strongly on the retained prefix rather than solving the problem independently of it; this does not exclude further computation from the prefix, which the closed calls often perform, but it is why the stop's cost depends on what the prefix already contains.

B.2 Requests, prefills and the end-of-thinking token

What an answer request inside the open block produces. Two AMC23 16K stops show the typical invalid case. In the first, the retained prefix ends with "So the answer is 27."; the strict call inside the open block continues "Wait, the prompt says 'Give the final answer as a signed integer in \boxed{N}.' Usually, distance is non-negative. Is there any ambiguity?" and re-derives the meeting point in a new coordinate system until its 1K allowance runs out mid-equation, whereas the same call after the closure package returns `\boxed{27}` in eight tokens. In the second, the prefix has derived $x = 3y$; the open call begins "Wait, I should check the prompt requirements again. […] Okay, 36 is

Table 12: **With another problem's prefix, the closure readouts fail and often submit that problem's answer.** Accuracy (%) at AC's fired stops of the control cells with the stop's own prefix, with the AC stop prefix of another problem (pointwise 95% interval) and on the empty trace, and the share of stops (%) at which the submission is the other problem's answer and wrong for this one; number of stops in parentheses.

Cell	Interface	Own prefix	Other problem's prefix	Empty trace	Other problem's answer
AMC23 16K (169)	strict closed	87.0	7.1 [1.1, 15.2]	17.8	75.7 [65.6, 84.9]
	native closure	71.6	7.1 [0.6, 15.3]	45.0	61.5 [50.9, 72.1]
	staged cascade	44.4	3.0 [0.0, 6.6]	50.9	23.7 [16.2, 31.7]
SMT25 32K (194)	strict closed	60.8	0.0 [0.0, 0.0]	5.2	56.7 [48.4, 65.1]
	native closure	30.4	0.5 [0.0, 1.6]	0.5	26.8 [19.3, 34.7]
	staged cascade	6.2	1.5 [0.0, 4.1]	8.2	3.1 [0.6, 6.1]
AMC23 32K (183)	strict closed	87.4	6.6 [1.1, 14.1]	16.4	76.5 [66.8, 85.4]
	native closure	68.3	6.0 [0.5, 13.6]	41.5	59.6 [49.4, 69.5]
	staged cascade	46.4	3.3 [0.5, 6.8]	47.0	25.1 [17.8, 33.0]

an integer. Let's double check if I misread the question.", re-derives the solution and again exhausts its allowance, and the closed call returns `\boxed{36}`. Inside the open block the answer request is read as more material to reason about; on AMC23 16K the strict call there exhausts its allowance on 91% of the 200 trajectories under AC and after closure on 3%, and Appendix B.4 gives the resulting shares of invalid submissions.

A late check at the token level shows that after the strict request inside the open block the end-of-thinking token has a mean next-token probability of 0.006 to 0.06% at AC's stops and at the open endpoints of the control cells, so the model does not take the request as a cue to end its reasoning; the first generated token is a newline in both block states, a single one inside the block and a double one after closure.

Wording of the request. A follow-up repeats the strict pair with two further wordings at AC's stops and the source endpoints of the control cells: an external directive ("Stop thinking now. Write only your final answer, in \boxed{}.") and a sentence in the model's own voice ("Okay, I have spent enough time on this. I will now write my final answer in \boxed{}."). The failure inside the open block does not depend on the wording: on AC's eligible stops the request inside the block is correct on 3.6 to 12.1% and exhausts its allowance on 98.8 to 100% under every wording, and closing the block adds 34.7 to 77.6 points, every interval excluding zero. The task's own instruction gives the most accurate closed answer; after closure the two new wordings more often reason at length before answering.

Budget sentence and delimiter. The 25-token closure splits exactly into a sentence segment (a blank line, the 21-token budget sentence and a newline) and a delimiter segment (the end-of-thinking token and a blank line). Four strict one-call arms share the prefix, the strict final-answer instruction, greedy decoding and a 1,024-token allowance and differ only in which segments precede the instruction: none (strict open), sentence only, delimiter only, both (strict closed). Injected tokens are outside the allowance. The primary family has nine contrasts: the sentence main effect, the delimiter main effect and their interaction on AC-eligible stops of AMC23 16K, SMT25 32K and AMC23 32K. Figure 8 shows the primary family, with a simultaneous half-width of 11.2 points, next to the same effects in six further Qwen cells and on DeepSeek-R1-Distill-Llama-8B: for the strict request the delimiter carries the effect in every Qwen cell, the budget sentence stays within a few points of zero, and on DeepSeek neither the delimiter nor the interaction is detected. Secondary contrasts agree. At designated-EAT stops the delimiter effect is 67.0, 49.2 and 79.5 points and the sentence effect lies between -1.7 and $+2.3$. On eligible horizon-cut endpoints the delimiter effect is 57.5 points on AMC23 16K and 28.7 on SMT25, smaller than at AC stops by 13.4 ($[-0.6, 28.1]$) and 18.9 ($[6.3, 31.1]$) points; AMC23 32K has 15 eligible endpoints and is not estimated. Allowance exhaustion on eligible AC stops is 97–99% for strict open, 91–96% with the sentence alone, 8–42% with the delimiter alone and 3–34% with both. The four arms coincide on all 585 ineligible stops.

	Strict call on AC-eligible stops, accuracy (%)				Native closure	Effect (pp)	
	Neither	Sentence	Delimiter	Both		sentence	delimiter
sentence	○	●	○	●			
delimiter	○	○	●	●			
Qwen3.5-9B control cells							
AMC23 16K (187)	9.1	13.4	80.7	83.4	63.1	+3.5	**+70.9**
SMT25 32K (246)	4.1	6.1	49.2	56.1	26.0	+4.5	**+47.6**
AMC23 32K (180)	11.7	10.6	84.4	86.7	65.0	+0.6	**+74.4**
Further Qwen cells							
AIME25 16K (145)	1.4	0.0	32.4	36.6	17.2	+1.4	**+33.8**
AIME25 32K (140)	4.3	3.6	47.9	40.0	18.6	**−4.3**	**+40.0**
BRUMO25 32K (140)	2.9	6.4	58.6	58.6	33.6	+1.8	**+53.9**
HMMT-Feb25 32K (147)	5.4	4.1	46.9	47.6	15.0	−0.3	**+42.5**
CMIMC25 32K (188)	2.7	5.9	30.9	30.3	11.7	+1.3	**+26.3**
AMC23 16K, 27B (190)	16.3	12.1	82.1	90.0	62.6	+1.8	**+71.8**
DeepSeek-R1-Distill-Llama-8B							
AMC23 16K (176)	40.9	47.7	40.3	50.6	38.1	+8.5	+1.1

Figure 8: **For the strict request the end-of-thinking token carries the closure effect; the budget sentence does not.** Accuracy (%) on AC-eligible stops (their number in parentheses) of the strict call at 1K with neither segment, the budget sentence only, the delimiter (end-of-thinking token) only, or both, and of native closure's free answer; darker cells are more accurate, and the dots under each column head mark which segments precede the instruction (filled: present). Effects in points; bold: the interval excludes zero (the simultaneous band of the experiment's family for the Qwen control cells, half-width 11.2, and for DeepSeek-R1-Distill-Llama-8B; pointwise for the further cells). For DeepSeek the delimiter is its own end-of-thinking token and the sentence is Qwen's, transferred.

Strict pair and factorial on further cells. A follow-up runs the four strict arms of the factorial at 1K on the source-endpoint, designated-EAT and AC stops of six further Qwen cells: AIME25 at 16K and 32K; BRUMO25, HMMT-Feb25 and CMIMC25 at 32K; and AMC23 16K on Qwen3.5-27B. Strict closed exceeds strict open by 27.7 to 73.7 points on eligible AC stops, every contrast excluding zero in the six-contrast simultaneous band (half-width 16.2). The delimiter carries the effect in every cell, and composition carries the larger share of the operational interaction in five of the six, AIME25 16K being the exception (Figures 7 and 8). On the 560 ineligible stops the four arms coincide.

A prefilled answer inside the open block. A final follow-up, specified after all earlier results, asks whether the end-of-thinking token matters when the answer is not requested but begun. On the control stops it crosses two cues with the block state: the crossing's trial cue ("The final answer is \boxed{") and the probe of Dynasor's released code, each with or without the end-of-thinking token before it, in one greedy request of at most 48 tokens read like the stored trial. With the cue alone the model almost always completes the box inside the open block, and the answer is about as accurate as after closure (Table 13); the six delimiter effects at AC stops lie between -2.1 and $+0.6$ points, none excluding zero in a simultaneous band of half-width 2.8, against $+47.6$ to $+74.4$ for an instructed request. At the designated EAT stops and at the open horizon-cut source endpoints the delimiter effects lie between -2.8 and 0.0 points. The regenerated closed request reproduces the stored trial on 923 of 926 rule stops. The end-of-thinking token is therefore needed for an answer that is requested and may still reason, not for one whose opening is forced; decoded greedily, Dynasor's open-block probe reads these stops about as well as its closed counterpart. The released probe is sampled at temperature 0.6 for 20 tokens; here decoding and length follow the crossing's trial so that only cue and block state vary. The instructed request and the prefilled call also differ in instruction and allowance. Every prefilled call ended at its stop string within 48 tokens, so a 1K allowance returns the same answers; truncating the greedy strict calls to their first 48 tokens gives the request at a 48-token allowance. At a common 48 tokens, closing the block raises the request's accuracy at AC's stops by 34.9, 16.0 and 38.8 points (from 4.7, 6.7 and 4.9%; every interval excluding zero) and the prefilled answer's by $+0.6$, -2.1 and $+0.5$, so the dependence on the block state belongs to the answer mode, not to the allowance; with 48 tokens a closed request often has not yet written its box.

Table 13: With the answer's opening prefilled, no delimiter effect is detected. Accuracy (%) at AC's stops (the n fired trajectories) of one greedy request of at most 48 tokens that begins the boxed answer, after the end-of-thinking token (closed) or inside the open block (open), with the crossing's trial cue or Dynasor's released probe; the delimiter effect is closed minus open (pp) with the simultaneous band of the six-contrast family.

		Trial cue		Dynasor probe		Delimiter effect	
Cell	n	Closed	Open	Closed	Open	Trial cue	Dynasor probe
AMC23 16K	169	88.8	88.2	88.2	87.6	$+0.6$ [-2.2, 3.4]	$+0.6$ [-2.2, 3.4]
SMT25 32K	194	71.1	73.2	72.2	72.7	-2.1 [-4.8, 0.7]	-0.5 [-3.3, 2.3]
AMC23 32K	183	88.5	88.0	88.0	87.4	$+0.5$ [-2.2, 3.3]	$+0.5$ [-2.2, 3.3]

Why the stored trial wins at 1K. On the AC stops where the stored trial is correct and native closure is not, closure exhausted its allowance without a legal answer in 493 of 501 cases across models (all 432 cases of the nine Qwen3.5-9B crossing cells and all 51 Qwen3.5-27B cases); every case in the other direction is closure correcting a wrong trial. Ministral-3-8B is the only model with closure answers that change a correct candidate.

Stored and fresh boxed trials. Rebuilding the trial request exactly at the AC, designated-EAT and last horizon-cut boundaries of the three primary cells reproduces the stored candidate and token count on all 1,071 stop boundaries. Regrouping the requests into different batches, which measures the inference engine's own nondeterminism, changes 1.2% of the candidates ($[0.6, 2.0]$) and correctness on two stops, so stored-trial accuracy is unchanged by regeneration.

B.3 Answer allowance

Answer allowance with the source fixed. Continuation, native closure, strict open and strict closed receive 2K and 4K on the source endpoint and on the designated-EAT and AC stops of the three control cells, with the source unchanged, so the total may exceed the original cap: a mechanism diagnostic, not an equal-cost comparison. Native closure keeps its injected tokens inside its allowance, the strict arms outside; all $3{,}596$ continuation and native-closure answers already generated on stops with room inside the cap are reused and match exactly. The primary family has twelve contrasts, τ_{AC} for strict closed minus strict open and minus native closure in three cells at 2K and 4K (simultaneous half-width 12.4), and the two closed readouts count as equivalent only if the 90% interval of their equal-cell pooled difference lies within ± 5 points (Table 14).

Two results follow. First, the closed readouts converge once the answer has room: strict closed minus native closure, pooled over the cells, is $+24.0$ points at 1K but -0.6 at 2K and -1.9 at 4K, within ± 5 at both, also with task-equal weights. Second, the open block does not: a strict call inside it still exhausts its 4K allowance on 89–94% of eligible AC stops, against at most 19% after closure, and closure helps AC's stops more than eligible horizon cuts by 27 to 34 points at 2K and 4K; at the later designated-EAT stops the open-block deficit shrinks from 58–69 points at 2K to 32–52 at 4K. Under the common strict closed readout the rule comparisons of Section 4 shrink (Appendix B.4).

Strict readouts for the remaining stops. Two further follow-ups read the stops that Figures 2 and 4 would otherwise leave without the strict pair: the random cuts and the empty trace at 2K and 4K, the s1-style forcing stops, and the matched EAT stops of AIME25 16K, with the settings of the answer-allowance experiment. The random cuts behave like the rule stops they imitate. With an empty trace the strict call after closure does not change with the allowance, whereas the same call inside the open block, where the model may still reason, can gain (to 26.8% on SMT25 at 4K).

Table 14: More room brings the two closed readouts together but leaves the open block far behind. Pooled over the cells, strict closed minus native closure is $+24.0$ points at 1K, -0.6 at 2K (90% interval $[-3.2, 1.9]$) and -1.9 at 4K. Accuracy (%) on the sources eligible under AC (block open and no answer: the prefix at AC's stop, or the source endpoint where AC does not fire; 187, 246 and 180) under continuation, native closure, strict open and strict closed; τ contrasts on those stops and the operational interaction I with the source endpoint for strict closed minus strict open, in points. 1K rows are the controls' arms; †: excludes zero in the simultaneous band over twelve 2K/4K contrasts.

		Accuracy on eligible AC stops (%)				τ: strict closed minus		
Cell	R	Cont.	Native	Open	Closed	open	native	I
AMC23 16K	1K	3.7	63.1	9.1	83.4	$+74.3$	$+20.3$	$+45.5$
	2K	4.8	88.2	16.0	87.2	$+71.1^{\dagger}$	-1.1	$+49.0$
	4K	20.3	89.8	20.3	87.2	$+66.8^{\dagger}$	-2.7	$+49.5$
SMT25 32K	1K	3.7	26.0	4.1	56.1	$+52.0$	$+30.1$	$+38.1$
	2K	4.5	67.9	8.1	69.5	$+61.4^{\dagger}$	$+1.6$	$+46.4$
	4K	12.2	69.9	13.8	70.7	$+56.9^{\dagger}$	$+0.8$	$+44.5$
AMC23 32K	1K	5.6	65.0	11.7	86.7	$+75.0$	$+21.7$	$+63.0$
	2K	10.6	91.1	13.3	88.9	$+75.6^{\dagger}$	-2.2	$+63.5$
	4K	20.0	93.3	20.0	89.4	$+69.4^{\dagger}$	-3.9	$+59.0$

B.4 Rule comparisons and the shared readout

Paired system comparisons. This subsection gathers every comparison between the two rules, each under its own procedure (AC's stored trial answer; for EAT our native-closure approximation, and below EAT's suffix with an 8K ceiling) and both under one shared readout. Under their own procedures at 1K, AC minus the EAT threshold of similar saving is $+18.0$ to $+36.2$ points in the three primary cells and at two further AMC23 caps; under any one shared generating interface or the stored trial it lies between -8.0 and $+7.3$. Table 15 adds the 2K allowance and the common strict closed readout.

Under the common strict closed readout. With the arms of the answer-allowance experiment (Appendix B.3), the comparisons of Section 4 shrink under the common strict closed readout: AC minus the matched EAT is -4.0 ($[-8.5, 0.0]$), $+3.4$ ($[-1.9, 9.1]$) and -2.5 ($[-7.0, 2.0]$) at 1K on AMC23 16K, SMT25 32K and AMC23 32K and stays within that range at 2K, all intervals including zero (on AIME25 16K it is $+5.3$, $[0.7, 10.0]$, under a readout that fails the validity check), and AC at 16K minus the AMC23 8K endpoint is -1.0 and 0.0, against -15.0 and -10.5 under strict open. The matched threshold, 2^{-1}, stops almost every trajectory it cuts at the end of EAT's warm-up and so acts as a fixed-boundary cut (Appendix A.1).

EAT's closure at 1K and 2K. A follow-up reads the matched-saving EAT stops of AMC23 16K, SMT25 32K and AIME25 16K with native closure at a 2K allowance and compares each rule under its own procedure on identical trajectories. AC's 1K lead becomes -7.5 ($[-16.5, 1.0]$), -5.3 ($[-11.3, 1.1]$) and 0.0 ($[-5.3, 5.3]$) at 2K, because the larger allowance cuts native closure's allowance exhaustion on the fired stops from 31–75% to 6–22%. The change itself is -26.5 ($[-37.0, -16.5]$), -41.5 ($[-50.6, -32.5]$) and -36.0 ($[-50.0, -22.7]$) points, each interval excluding zero.

EAT's answer suffix with an 8K ceiling. The EAT paper appends the end-of-thinking token and "Final answer:" and generates to EOS without fixing a capped allowance. Our follow-up reads the designated and matched EAT stops and the open, horizon-cut source endpoints of the three primary cells this way (greedy, at most 8,192 new tokens, so the answer may exceed the cap) and compares each rule under its own procedure (Figure 9, first three rows). With this 8K ceiling instead of 1K native closure, EAT's stop cost at the matched threshold falls from 10.0, 41.9 and 11.3 points to 2.0, 11.7 and 5.3 on AMC23, SMT25 and AIME25, and AC's 1K lead becomes -7.0, -3.4 and $+2.0$; at the designated threshold EAT is ahead on AMC23 (-10.0, $[-19.5, -1.5]$) and SMT25 (-13.6, $[-19.6, -7.5]$). A second follow-up repeats this in the six other Qwen3.5-9B crossing cells. The

Table 15: **Read with procedure-specific interfaces at 1K, AC leads the matched EAT threshold by 19–36 points; with 2K, or under a common readout where that readout is valid, the gap is no longer detected.** AC−EAT accuracy (pp) with pointwise paired 95% intervals; positive values favor AC. Own interfaces: AC's stored trial against EAT's native closure at a 1K or 2K allowance. Common: one strict call after the closure package for both rules; –: the 2K common readout was not run on AIME25 16K. †: this readout submits a valid answer on fewer than half of AIME25 16K's eligible stops (Appendix B.4). EAT's published procedure, which generates its answer to the end of the sequence, fixes no allowance.

	Reasoning saved (%)		Own interfaces		Common strict closed	
Cell	AC	EAT	1K	2K	1K	2K
AMC23 16K	51.5	47.3	+19.0 [8.0, 30.0]	−7.5 [−16.5, 1.0]	−4.0 [−8.5, 0.0]	−2.5 [−6.0, 0.5]
SMT25 32K	32.7	31.1	+36.2 [26.0, 46.4]	−5.3 [−11.3, 1.1]	+3.4 [−1.9, 9.1]	+3.4 [−1.1, 8.3]
AIME25 16K	25.9	25.5	+36.0 [22.7, 49.3]	0.0 [−5.3, 5.3]	+5.3 [0.7, 10.0]†	–

pattern holds in every one: AC's lead of 18 to 39 points under 1K native closure becomes −6.7 to +5.0 under this EAT-style readout, and at the designated threshold EAT is ahead by 8.0 to 12.7 points in five of the six. On MATH-500 level 5, specified before its generation, the same procedure turns AC's leads of 28.2 and 18.6 points into −1.4 ([−4.7, 1.6]) and −4.1 ([−7.6, −0.9]) (last row). Across the ten cells and both thresholds the difference moves toward EAT by 18.5 to 46.7 points, every pointwise interval excluding zero, and no cell leaves AC ahead with an interval excluding zero. The answers are long: a median of 520 to 1,080 tokens (6,660 at CMIMC25's matched threshold), more than 1K on up to 63% of the stops and at the ceiling on up to 48%, where an unfinished answer may be scored wrong. Which procedure-specific comparison one reports therefore depends on an allowance the EAT paper does not set, and neither comparison holds answer cost fixed: AC's trial is at most 48 tokens.

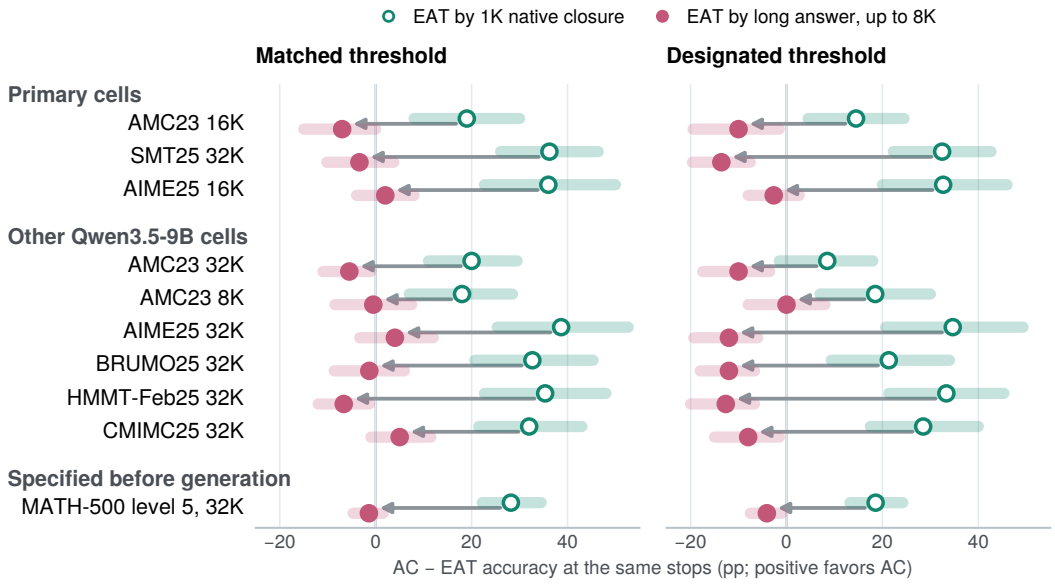

Figure 9: **An EAT-style long answer removes the detected AC lead in all ten cells.** AC, read by its stored trial (at most 48 tokens), minus EAT read by 1K native closure (open) or by the paper's suffix with a greedy answer of up to 8K tokens (filled), at the same stops; shaded bars are pointwise paired 95% intervals. The comparison changes EAT's suffix and allowance together and does not match the rules' answer costs.

Does the common readout submit an answer? The checklist of Section 8 asks for a shared readout whose valid-answer rate reaches 95% on the full traces and on every rule's stops. At 1K, strict closed

meets this on AMC23 but not on SMT25, where under the legal-box rule it leaves 14.0% of source endpoints and 31.3% of AC stops without a valid answer, against 1.0 and 3.0% on AMC23 16K and 0.0 and 2.5% on AMC23 32K; inside the open block the same call leaves 79.5 to 90.6% of AC stops without one. Among the eligible sources of cells opened later, it is valid on 96.3–100% for Qwen3.5-27B on AMC23 and on about 90% on GPQA-Diamond (so on at most 94.8% of AC's stops, counting the stops that already hold an answer), but on only 36.0–44.1% on AIME25 16K and 51.4–61.0% on AIME25 32K; on DeepSeek-R1-Distill-Llama-8B the model's own closure followed by the strict call is valid on 52.8–54.5%, and inserting Qwen's budget sentence raises this to 90.9–92.0%. By the checklist's own criterion, a 1K strict closed readout is thus an adequate shared readout on AMC23, just short of it on GPQA-Diamond, and inadequate on SMT25 or AIME25, where the allowance must be raised or the readout adapted, so the AIME25 comparison under this readout is not interpreted.

Prospective readout screening and the run-time acceptance audit. Before generation on MATH-500 level 5 and Ministral, we fixed a selection rule: choose the smallest shared strict-closed allowance among 1K, 2K and 4K reaching 95% valid submissions at every endpoint type. The MATH-500 implementation also counted source candidates accepted by its entry reader. That reader accepted the literal placeholder "..." from quoted answer-format instructions and skipped generation at 18 AC, 22 matched-EAT and 28 designated-EAT stops (67 distinct prefix keys). These outputs are invalid under the legal-box criterion. Rescoring the recorded outputs gives 2K rates of 98.27% (full), 94.17% (designated EAT), 93.86% (matched EAT) and 92.91% (AC); at 4K the last three are 94.17%, 93.86% and 93.70%. No recorded arm passes the screen (Figure 10). These rates describe the historical execution; validity after executing the skipped requests remains undetermined. The same source-candidate acceptance skips the cascade at these 18 AC stops, whereas continuation and native closure generate. No placeholder occurs among the 30,663 stored AC trial candidates in this cell, and no full endpoint has a placeholder source candidate.

The recorded 2K arm is reported as a diagnostic of that implementation. Under it AC minus matched EAT is 0.0 ($[-2.0, 2.0]$) and AC minus designated EAT is -0.8 ($[-3.5, 1.7]$), versus $+28.2$ ($[22.2, 34.5]$) and $+18.6$ ($[13.2, 24.3]$) under procedure-specific 1K readouts. At 2K the invalid shares are 7.09% for AC and 6.14% for matched EAT; if only those submissions could be repaired, with all valid outcomes held fixed, the observed zero gap could range from -6.14 to $+7.09$ points, a worst-case sensitivity range rather than a confidence interval. The paired interval conditions on the 2K arm selected from these evaluation trajectories and excludes uncertainty in that selection. Both rules have a 5.2-point stop cost under that request, but neither this equality nor the screen establishes an interface-independent ranking.

On Ministral-3-8B (AMC23 and AIME25, 16K) no allowance qualifies. After the transferred closure package and the strict instruction the model ends its answer at once (a single token at every AC stop), so the closed call is valid on at most 7.5% of AC's stops at every allowance. The checklist therefore reports no comparison under a shared readout for this model; under the procedure-specific 1K readouts AC minus the matched EAT (2^{-3} and 2^{-5}) is -8.5 ($[-19.0, 1.0]$) and -4.7 ($[-14.0, 4.7]$), and AC minus the designated EAT is -29.5 ($[-41.0, -19.0]$) and -10.7 ($[-22.0, 0.7]$). A late check repeats the rule with the model's own end-of-thinking token alone and adds a third-party candidate, the released Dynasor prefill probe inside the open block (48 tokens, greedy). The own delimiter changes almost nothing (median answer one token), so removing the transferred sentence does not resolve the early end. This is an interface failure under the implemented template; the prefilled request does produce answers. The prefilled readout clears the level at every stop type on AMC23 and is selected there: AC minus the matched EAT becomes $+7.5$ ($[1.0, 15.0]$), reversing the sign of the observed difference, and AC minus the designated EAT is -36.5. On AIME25 it does not (lowest rate 76.0%), and no readout qualifies. A prefilled readout resembles AC's own probe, so the comparison remains conditional on a readout similar to AC's selection probe.

Valid-answer rate (%) of each candidate readout

	Full trace	EAT, designated	EAT, matched	AC	at least 95% at all four
Qwen3.5-9B, MATH-500 level 5					
Open request, 1K	94.3	48.4	14.5	13.2	
Closed request, 1K	95.9	87.7	82.4	80.9	
Closed request, 2K	98.3	94.2	93.9	92.9	
Closed request, 4K	98.3	94.2	93.9	93.7	
Ministral-3-8B, AMC23 16K					
Open request, 1K	86.0	86.5	75.5	75.0	
Transferred closure, 1–4K	61.5	42.5	10.0	7.5	
Own delimiter, 1–4K	61.5	42.5	10.0	7.5	
Third-party prefill, 48	97.5	97.5	97.5	98.5	**selected**
Ministral-3-8B, AIME25 16K					
Open request, 1K	56.7	54.7	54.0	59.3	
Transferred closure, 1–4K	26.7	6.0	0.7	0.0	
Own delimiter, 1–4K	26.7	6.0	0.7	0.7	
Third-party prefill, 48	82.0	76.0	79.3	86.0	

Figure 10: **Recorded MATH-500 outputs fall below the screening level; a third-party prefill qualifies on Ministral's AMC23.** Valid-answer rate (%) of each candidate at the full trace and at the stops of designated EAT, matched EAT and AC (unfired trajectories contribute their source endpoint). MATH-500 rows apply legal-box validity to historical outputs, including the incorrectly skipped requests described above. Dark cells reach the pre-specified 95% level, and a readout passes only when all four do. On Ministral-3-8B the separately tested third-party prefill passes on AMC23; none qualifies on AIME25. Passing this screen alone does not establish a neutral or accurate readout.

B.5 Scope and cost

The published AC stopping definition, on one of its models. A follow-up applies the sentence-end Answer Convergence stopping definition ($k = 10$) to our one-shot DeepSeek-R1-Distill-Llama-8B source trajectories on AMC23 16K (200 trajectories) and AIME24 16K (30 problems, one seed). It crosses our terminal interfaces rather than reproducing the original end-to-end protocol. The AC paper describes greedy trial answers after the end-of-thinking token; its current public collection script (`https://raw.githubusercontent.com/launchnlp/reasoning_earlystop/main/src/collect_partial_reasoning_res.py`, checked September 2026) additionally prefills `\boxed` and forces EOS after a generated newline. Our free native closure does neither, and our 48-token boxed trial is also not that script's exact request. The rule fires on 183 and 30 trajectories at a median depth of 38 and 28% and saves 66 and 72% of reasoning tokens. With our free closure the stop costs 48.0 points on AMC23 ($[36.0, 60.0]$) and 23.3 on AIME24 (Table 16); on AMC23 the stored trial and strict closed request each cost 35.5 points, 12.5 less than free closure for the trial ($[3.0, 23.0]$). The strict-pair interaction is $+9.0$ ($[4.0, 14.5]$), and the cascade costs 8.5 points ($[1.0, 16.5]$) less than closure, the opposite of Qwen3.5; closure against continuation, $+2.5$ ($[-3.0, 8.5]$), is not detected. On AIME24 the free-closure accuracies, 40.0% at full traces and 16.7% at stops, numerically match the AC paper's baseline and $k = 10$ rows for this model, but this numerical match does not establish protocol equivalence. That paper also reports a 23.0-point loss on its MATH-500 split (79.0 to 56.0%) and does not evaluate AMC23. On the 30 AIME24 problems our readouts stay within ten points of one another and no interaction is detected.

One-shot sources. The crossing is repeated on AMC23 16K sources generated by one request to $B - 1{,}024$ with the same sampling parameters and seeds, so the presence penalty is never reset; strict open and strict closed are added at 1K. The source endpoint reproduces the source bank's earlier one-shot outputs on all 200 trajectories, the outcome-blind rule again designates $\delta = 2^{-9}$, and firing and saving resemble the chunked sources'. In the six-contrast family (half-width 15.5), native closure minus continuation ($+38.0$ for AC, $+21.5$ for EAT) and strict closed minus strict open on eligible stops ($+72.2$, $+72.1$) exclude zero, whereas cascade minus closure (-11.0, -5.0) does not; composition carries 40 of the 41-point operational interaction.

A non-mathematical crossing. GPQA-Diamond (Rein et al., 2023) (198 questions) is crossed at 16K on new chunked sources of Qwen3.5-9B, read by the official GPQA reader, which takes the

Table 16: **With sentence-end AC stops, our terminal interfaces move the measured cost by up to 15 points.** DeepSeek-R1-Distill-Llama-8B on our source trajectories: full and stop accuracy (%) and stop cost (pp), with pointwise 95% problem-cluster intervals. Free native closure is our diagnostic arm, not a verified reproduction of AC's published answer request.

Cell	Interface	Full trace	AC	Stop cost
AMC23 16K	open continuation	87.0	36.5	+50.5 [39.0, 62.0]
	free native closure (ours)	87.5	39.5	+48.0 [36.0, 60.0]
	staged cascade	87.5	48.0	+39.5 [28.5, 51.0]
	strict open	87.5	43.0	+44.5 [32.5, 56.5]
	strict closed	87.0	51.5	+35.5 [24.0, 47.5]
	stored trial (prefilled)	87.5	52.0	+35.5 [24.5, 47.0]
AIME24 16K	open continuation	40.0	13.3	+26.7 [13.3, 43.3]
	free native closure (ours)	40.0	16.7	+23.3 [10.0, 40.0]
	staged cascade	40.0	16.7	+23.3 [10.0, 40.0]
	strict open	40.0	10.0	+30.0 [13.3, 46.7]
	strict closed	40.0	20.0	+20.0 [6.7, 33.3]
	stored trial (prefilled)	40.0	16.7	+23.3 [10.0, 40.0]

first parenthesized letter in the text, with AC's trial cue adapted to "The correct answer is ("; the eleven-contrast family adds the strict pair to the nine interactions (half-width 7.4). The outcome-blind rule designates $\delta = 2^{-3}$ (14.9% saving; 535 fired trajectories); AC fires on 690 and saves 29.0%. Closure minus continuation and cascade minus closure exclude zero for both rules (+19.9 and −18.0 for EAT, +24.3 and −21.8 for AC), as does the strict pair on eligible stops (+55.5 and +52.9), whereas cascade minus continuation and all three forcing-control contrasts do not. Unlike on AMC23, the strict pair works through both terms, 33.8 points of validity and 19.1 of conditional correctness for AC, and eligible early stops respond more than eligible endpoints ($\tau_{\text{AC}} - \tau_{\text{Full}} = +7.2$, $[0.3, 14.2]$); composition still carries most of the operational interaction (about 13 of 16.2 points). The reader shapes the costs as well: because it takes the first letter anywhere, 72% of source endpoints already hold an answer, and the stored trial, which submits one letter after closure, reaches 71.7% on the full traces against 59.9% for native closure, so that early stops cost nothing under it (AC -0.7, $[-3.5, 2.2]$; EAT -3.9, $[-6.0, -1.9]$). Because the reader takes the first letter anywhere, a later check rescores the same submissions with the reader applied only to the answer region (the text after the last end-of-thinking token, else the generated text) and, as a candidate-only reading, to the last letter in that region. AC's closure-versus-continuation interaction is +19.9 ($[14.9, 25.1]$) and +12.9 ($[8.3, 17.7]$) instead of +24.3, EAT's +16.9 and +13.8 instead of +19.9, and the strict pair's contrast on AC's eligible stops +43.9 and +27.2 instead of +52.9, every interval excluding zero.

A second non-Qwen family. DeepSeek-R1-Distill-Llama-8B is crossed on AMC23 16K with one-shot sources, an EAT threshold frozen on AIME24 before any AMC23 source was generated, and its own end-of-thinking token as the closure; Qwen's budget sentence enters only as a factor. The model thinks briefly (187 of 200 traces end on their own), so only 5.5% of source endpoints are eligible. The frozen threshold (2^{-7}) fires on 10 trajectories and EAT is not interpreted; AC fires on 175 and saves 67.2% of reasoning tokens. None of the six primary contrasts on eligible AC stops excludes zero in its band (half-width 13.3), including strict closed minus strict open (-0.6 at 1K and at 4K) and the delimiter effect (+1.1); the sentence effect (+8.5) excludes zero only pointwise. The difference from Qwen3.5-9B lies in the open block: a strict call inside it is already valid on 64.8% of eligible stops, and adding the end-of-thinking token trades validity (-8.3) for conditional correctness (+7.8). The answer procedure still matters: Qwen's sentence before the model's own delimiter, followed by a free answer, reaches 59.1% on eligible AC stops, 21.0 points above the model's own free closure, and native closure lowers AC's cost relative to continuation by 8.5 points, carried by validity as in every crossing cell (Table 10).

Serving cost. Probes and trial answers are sampled output charged outside B, but under a black-box API each of them also re-reads its prefix. Table 17 reports tokens per trajectory for the AMC23 cap comparison under three regimes: a persistent KV fork (decoding only, prompt read once), a prefix cache (decoding plus uncached suffixes) and a cold black box (every call re-reads its full input). The accuracy–cost status of AC at 16K against the 8K endpoint changes between regimes for all five

interfaces, whereas against the fixed-token cut it does not. Designated EAT, which probes at every checked boundary, costs 11 times the 16K full trace under a cold black box. These are token counts, not FLOPs or latency.

Table 17: Whether AC at 16K is cheaper than an 8K cap depends on the serving regime: under native closure it uses 0.89 times the 8K cap's tokens with a persistent KV cache and 4.2 times as many when every call re-reads its prefix. AMC23: AC and designated EAT at 16K, the full traces at 16K and 8K, and the fixed-token cut matched to AC's mean reasoning length, under each interface (the Parser is the offline diagnostic reader) and three serving regimes (tokens per trajectory, prefill and decode weighted equally; the last column weights a prefilled token a quarter of a decoded one).

Interface	Source	Acc. (%)	KV fork	Prefix cache	Cold	Cold, w=.25
Continuation	AC, 16K	8.0	7,204	7,354	60,108	20,302
	EAT, 16K	32.5	12,174	12,254	240,024	69,008
	Full, 16K	59.5	13,351	13,351	21,432	15,243
	Full, 8K	20.5	7,991	7,991	14,452	9,478
	Fixed cut, 16K	19.5	7,096	7,096	12,670	8,361
Native closure	AC, 16K	65.5	6,988	7,162	59,915	20,092
	EAT, 16K	67.0	12,031	12,130	239,901	68,870
	Full, 16K	72.5	13,308	13,318	21,400	15,202
	Full, 8K	57.0	7,879	7,900	14,362	9,372
	Fixed cut, 16K	57.5	6,967	6,989	12,563	8,238
Staged cascade	AC, 16K	45.0	6,882	7,053	62,550	20,671
	EAT, 16K	68.0	11,802	11,896	241,275	69,042
	Full, 16K	80.5	13,128	13,137	21,782	15,163
	Full, 8K	54.5	7,664	7,682	15,730	9,552
	Fixed cut, 16K	54.5	6,751	6,769	13,634	8,344
Stored trial	AC, 16K	81.5	6,208	6,358	53,220	17,833
	EAT, 16K	87.5	11,353	11,441	238,033	67,895
	Full, 16K	87.5	12,862	12,867	19,700	14,444
	Full, 8K	82.5	7,120	7,128	12,748	8,399
	Fixed cut, 16K	82.5	6,198	6,206	11,780	7,465
Parser (diagnostic)	AC, 16K	6.0	6,208	6,358	53,220	17,833
	EAT, 16K	25.0	11,350	11,430	229,558	65,774
	Full, 16K	52.5	12,860	12,860	12,860	12,732
	Full, 8K	17.0	7,116	7,116	7,116	6,988
	Fixed cut, 16K	17.0	6,195	6,195	6,195	6,066

C Validation and Interpretation

C.1 Pairing and replication gates

Pairing. In the crossing every policy is a stop point on, or a continuation of, one stored trajectory, and every interface replays the stored tokens of its stop point, so both directions of the interaction are paired and a never-fired trajectory contributes exactly zero. The bootstrap resamples problem identities and keeps the five seeds and every branch of a problem together.

Replication gates. Three exact-reproduction checks tie the runs together. In all five cells of the crossing the full-trace parser and cascade outputs reproduce the source bank's earlier outputs exactly, in submitted text, answer-stage tokens, token counts and correctness, on all 1,015 trajectories (200, 265, 150, 200 and 200); the AIME25 32K crossing cell reproduces the parser and cascade outputs of the existing AIME25 32K trajectories, 150 of 150; and an independent regeneration of the AIME25 16K parser and cascade arms reproduces them exactly, 150 of 150. These gates establish agreement between implementations of the specified protocol; the stop and submission definitions determine the quantity it measures.

Separation of the three readers. Diagnostic states are computed with gold answers, so they must not be able to reach a submission. An offline diagnostic Parser labels each source C, W or U and is

never consulted at run time. A run-time acceptance rule decides, without gold, whether a procedure skips generation. A final scorer evaluates the submitted string. On the mathematics benchmarks the readers take the most recent boxed candidate; on GPQA-Diamond they are the official ordered first-match patterns, so an early tentative letter outranks a later one (Appendix B.5).

Reproduction checks in the replication cells. Qwen3.5-27B and the held-out benchmarks use newly generated base trajectories from the same frozen sampling procedure; source-bank reproduction is therefore not applicable. The two Ministral-3-8B cells, whose generation procedure differs only where that model's end-of-thinking token, a control token absent from decoded text, requires it, reproduce the parser and cascade outputs of that model's earlier source bank exactly, 200 of 200 and 150 of 150 trajectories.

Limitations. The rules are adaptations: EAT follows its first arXiv version, whose matched threshold stops at the end of warm-up, and AC checks blank-line rather than sentence boundaries, so rule-versus-rule results describe these adapted operating points. The fixed-cut reversal is measured at a 1K answer allowance; under the stored trial no difference is detected between the rule and the fixed cut. The MATH-500 common request is a diagnostic comparison rather than a validated shared readout. Costs are token counts, not energy or latency, and the source bank and cascade come from earlier studies. Specifications are time-stamped internally rather than registered with a third party.

C.2 Answer procedures in published stopping methods, and related work

How stopped traces become answers in prior work. Published stopping methods use several different closures. EAT (Wang et al., 2025) and ThinkBrake (Song et al., 2025) close the thinking block and then generate freely; EAT appends "Final answer:" after the end-of-thinking token, generates to the end of the sequence without a fixed allowance, and in its experiments also computes its signal after a "Final answer: " prefix (Appendix A.1). Answer Convergence (Liu & Wang, 2025) returns its greedy, boxed trial answer after closing the block. DEER (Yang et al., 2025b) closes the block after a confident boxed trial answer and generates a greedy conclusion. s1 (Muennighoff et al., 2025) uses the end-of-thinking delimiter, optionally followed by "Final Answer:". Dynasor (Fu et al., 2025) and Conformal Thinking (Wang et al., 2026) force a final-answer cue and an open box. Dynasor's released code appends its cue, "… Oh, I suddenly got the answer to the whole problem, **Final Answer**" followed by an opening `\boxed{`, to the unfinished reasoning without an end-of-thinking token, samples a 20-token answer, and submits that answer when the probes agree; its answers are therefore formed inside the open block. At our stops such a cue, decoded greedily, is about as accurate inside the open block as after closure (Appendix B.2). The description of Conformal Thinking does not establish whether it ends the thinking block first. None of these seven papers reports the same stopping rule under two closures.

Two evaluation studies vary the answer step directly: Dong et al. (2026) compare three templates following a stop marker and a greedy answer of at most 48 tokens; one template reduces full-trace accuracy to 23%. Wang & Zhu (2026) compare free continuation with a forced boxed answer of at most 64 tokens; at 10% prefixes the former recovers 70% and the latter 34%. Their paper also leaves the thinking-block state before the forced cue unspecified. Apart from Dynasor, these descriptions come from the papers rather than their released code.

Section 7 gives the shorter account of the closest work; the remaining paragraphs place the closure comparison among related methods and measurement studies.

Budget control and stopping rules (the evaluated objects). Test-time compute can be scaled by repeated sampling or longer traces (Brown et al., 2024; Snell et al., 2024), steered toward a budget by prompting, guidance, or training (Han et al., 2025; Li et al., 2025; Aggarwal & Welleck, 2025; Sun et al., 2025), or cut short by a stopping rule. Stopping signals include repeated trial answers (Liu & Wang, 2025), entropy after the end-of-thinking token (Wang et al., 2025), trial-answer confidence (Yang et al., 2025b), certainty indices for serving (Fu et al., 2025), log-probability margins (Song et al., 2025), reflective redundancy (Sun et al., 2026), temporal aggregation (Li et al., 2026), semantic convergence (Min et al., 2026), answer-probability certainty (Nogueira et al., 2025), risk control (Wang et al., 2026), and hidden-state probes (Zhang et al., 2025). Each rule reports accuracy under its own answer step, which is the dependence this paper measures.

Answering from partial reasoning. Truncated traces can be answered directly (Liao et al., 2026), thinking can be skipped altogether (Ma et al., 2025), and deployed systems insert a stop-thinking instruction when a budget is exhausted (Yang et al., 2025a). Training-side versions give the answer its own budget: elastic reasoning separates thinking and solution phases (Xu et al., 2025), and anytime reasoning truncates thinking at sampled budgets and trains a summary policy (Qi et al., 2025). Interrupting a model mid-thought degrades accuracy and leaks reasoning into the answer (Wu et al., 2025), and forced extraction can fail where free continuation succeeds (Wang & Zhu, 2026). Our staged cascade is a composition of such known operations, used as a reference protocol rather than proposed as a new method.

Protocols and measurement validity. Budget-aware evaluation matches compute before comparing strategies (Wang et al., 2024a); evaluation protocols and unified benchmarks standardize prompts, extraction, and cost (Rumiantsev et al., 2025; Huang et al., 2025); a cost-aware study of early exits holds one answer step constant across stopping rules and reports that a brittle answer template can collapse accuracy (Dong et al., 2026); matched-horizon probes report completion and the stopped-prefix answer separately (Carlon et al., 2026); and answer regeneration removes dependence on extraction rules (Jo et al., 2025). Related evidence comes from outside reasoning control: learned extractors beat regular expressions (Yu et al., 2025), generated answers disagree with scored options (Wang et al., 2024b; Chandak et al., 2025), format restrictions cost accuracy (Tam et al., 2024), reasoning models violate format instructions (Kwon et al., 2025), output caps swing measured gaps (Goyal & Ray, 2026), and format repair can be mistaken for self-correction (Chen et al., 2026b; Huang et al., 2024). Truncation itself can be worse than no reasoning (Su et al., 2026), and budget and difficulty confound trajectory analyses unless a question-only or within-problem baseline is used (Bulut, 2026), which is why the controls of the crossing include an empty trace.

What the answer stage reads. Attention is biased toward recent context (Liu et al., 2024), but counterfactual importance concentrates on a few anchor sentences (Bogdan et al., 2025), and answer tokens re-read anchors as well as the tail (Chen et al., 2026a). Bounded-context reasoning keeps only a summary or the last tokens between chunks (Aghajohari et al., 2026; Yan et al., 2026).

Confounded comparisons and inference. Relative comparisons can depend on evaluation factors such as task-relevant pretraining (Dominguez-Olmedo et al., 2025), implementation and seed variance (Hochlehnert et al., 2025). Our statistical practice follows the recommendation to report interval estimates with the clustering the design implies (Miller, 2024; Agarwal et al., 2021).

Cost accounting. Answer-stage calls re-read the accumulated prefix, so their cost depends on the serving regime: prefix caching and KV reuse remove the prefill compute for a shared prefix but not the decoding of new tokens (Kwon et al., 2023; Zheng et al., 2024), black-box prefilling removes neither (Dong et al., 2026), and attention, not parameters, dominates test-time cost at long contexts (Sadhukhan et al., 2025). We therefore report sampled output, replayed input, and answer calls separately.

Anytime framing. A stopping rule plus an answer protocol is an anytime system: the quality available at an interruption depends on the contract for responding to it (Zilberstein, 1996). Read this way, the paper's claim is that an interruptible reasoner cannot be evaluated without specifying that contract.