CORRECTING IS NOT REUSING: WHAT A FAILED ATTEMPT CONTRIBUTES BEYOND THE VERDICT

Anonymous authors
Paper under double-blind review

ABSTRACT

A rejected answer leaves a retry two things, the verdict and the transcript that produced it, and self-correction scores them together. We separate them on exact-answer mathematics with a retry protocol in which both sides enforce the verdict. Both retries use the same output-token ceiling and a controller that discards every draw in the rejected answer class, with every solver draw billed. They differ in whether the failed transcript is retained or replaced by a withholding notice, with the instruction worded to match, and the difference in success is the retention contrast. In four instruction-tuned models not trained for retry it is negative: keeping the transcript lowers success by .106 to .235 (.235 on Llama-3.1-8B, 161 problems). The penalty persists after final-answer redaction: under one instruction at matched prompt length, the redacted transcript stays .074 to .260 below a content-free placeholder. With the transcript in context, the share of first calls that return to the rejected answer class rises by .17 to .50, and the controller's redraws recover only part of the loss. Among nine complete retry policies priced against a fresh draw, our full-history continuation baseline is the worst in every family and ceiling. LoRA supervised fine-tuning on retry prompts lifts the kept transcript above the withheld one on a fixed failure set in each of three seeds, yet end-to-end success stays below the untrained model's (.616 against .739, first seed). Under a one-step revision protocol, S2R's self-correction RL raises the correction rate while the revised retry still succeeds less often than a fresh draw from the same checkpoint. Correction rate, retention contrast and end-to-end success give different verdicts and must be reported separately.

1 CORRECTION IS NOT REUSE

A verifier rejects a large language model's answer y and the system tries again. Two different things are now available. One is new evidence: y is wrong. The other is old computation: the reasoning that led to y. Self-correction loops hand both back to the model (Madaan et al., 2023; Shinn et al., 2023; Kumar et al., 2025), and the success rate of such a retry credits the transcript with everything that follows: acting on the verdict, solving the problem again, and building on the earlier attempt. The first two need no transcript: a retry that cannot return to y does both without seeing the attempt. This paper asks what is left: *what does keeping the failed attempt add over the same retry with the attempt withheld, under the same verdict and the same output-token ceiling?*

We answer it with a retry protocol in which both sides enforce the verdict. Our headline comparison, the *retention contrast* J, is the success of a retry that keeps the failed transcript minus that of a retry in which the transcript is replaced by the line "[Previous solution withheld.]" and the instruction says that the solution is not shown; both run under one controller that discards every draw in the rejected answer class, with everything else shared and every solver draw billed (§2). Neither prompt says that the answer was wrong; the verdict acts only through the controller. J is what keeping the attempt adds beyond the verdict. Without the ban a kept retry could simply re-propose y, which the verdict already rules out; with it, what remains is budget spent on returns and new wrong answers. J is negative in all four models we test, released instruction-tuned checkpoints not trained for retry: keeping the transcript lowers success by .106 to .235 (Figure 1a). The two simple explanations, prompt length and the shown answer, do not account for it. A content-free placeholder of the tran-

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.

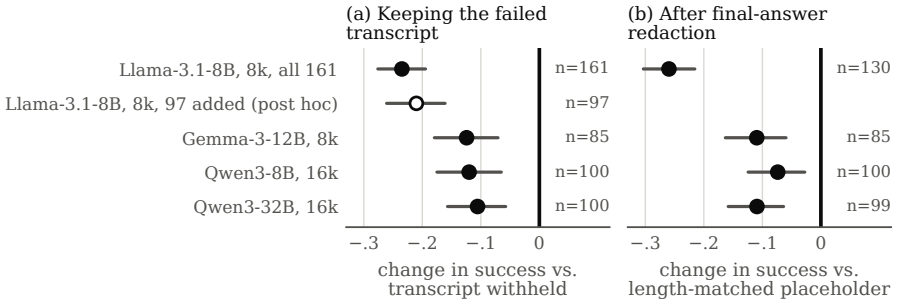

Figure 1: **Keeping the failed transcript lowers success in all four models on exact-answer mathematics, also with its final answer redacted.** Change in success with problem-paired bootstrap 95% intervals; n: problems. (a) The retention contrast J: the whole failed transcript kept, minus a retry with it replaced by a one-line withholding notice and the instruction worded to match. (b) The transcript with its final answer redacted against a content-free placeholder of matched prompt length, under one instruction, on the problems whose starts carry a boxed answer. Both panels: one controller that bans the rejected answer's class; Qwen3 with thinking off. In (a) the hollow marker is the 97 Llama-3.1-8B problems added after the 64-problem development set, a split made post hoc. Values: Appendix K.

script's length is not detected below the withheld retry in any model, and under one instruction at matched prompt length the loss persists after final-answer redaction: the redacted transcript stays below the placeholder in every model (Figure 1b). The transcript pulls the retry back: the share of first calls that return to the rejected class rises by .17 to .50, before the controller acts, and the controller's redraws recover only part of the loss (§4).

Two further collections place the contrast in context: nine complete retry policies, from nothing kept to the whole answer history, priced against a fresh draw under their own controllers (§5), and two kinds of retry training, supervised fine-tuning on retry prompts and S2R's self-correction RL (Ma et al., 2025), read on the correction rate, J and end-to-end success (§6).

Prior work compares the two ends, a restart and a retry that keeps the whole attempt, without enforcing the rejection on the restart. Olausson et al. (2024) match budgets between i.i.d. sampling and self-repair in code and find repair's gains often modest and sometimes absent; Yang (2026) proves that a clean restart dominates a contaminated retry whenever the contaminated error rate exceeds the clean one. Cheng et al. (2026) measure how a failed attempt in context pulls later generations toward similar errors across eleven models, and Verma (2026) finds blind resampling ahead of self-repair in small code models, whose repairs re-emit a near-identical program in 33–68% of retries against 2–14% under blind resampling. İşcan (2026) comes closest to holding the verdict fixed: a restart told only that its solution failed beats one that also sees the failed code, but in neither restart does a controller enforce the rejection. Partial retention has its own designs, a located correct prefix (Samanta et al., 2026; Wu et al., 2025b) and a kept chain with an error report (Chen & Li, 2024). Trained multi-turn correction reports gains over repeated sampling (Kumar et al., 2025; Qu et al., 2024; Liu et al., 2025), whose sampling baselines are not held to a rejection. Appendix A treats each line in full, with the matched-token comparisons of refinement against repeated sampling, the conditional correction statistic (Nilayam et al., 2026) and the reading of external verdicts as the source of self-correction gains (Huang et al., 2024; Stechly et al., 2024).

Contributions.

1. **A retry protocol that holds the verdict fixed on both sides** (§2). Kept and withheld retries differ in the retained turn and one instruction clause and share a controller that bans the rejected answer class, so their difference, the retention contrast J, measures what keeping the failed transcript adds beyond the verdict; a routing/proposal ledger splits any retry policy's gap to a fresh draw exactly (§4).
2. **Keeping the failed transcript lowers success, also with its final answer redacted** (§§3–4). J is negative in all four models (Figure 1a); under one instruction at matched prompt

length, the redacted transcript stays below a content-free placeholder (Figure 1b); and the transcript sends the first call back to the rejected answer, and on Llama-3.1-8B toward new wrong answers, a loss the controller's redraws recover only in part.

3. **Complete retry policies, each priced against a fresh draw** (§5). Nine policies that retain nothing up to the whole answer history, under their own instructions and controllers: our full-history continuation baseline is the worst in every family and ceiling and below a fresh draw in all but one further inference regime, and the verdict-only restart is not detected below a fresh draw in any setting where it ran.
4. **After training, correction rate, retention contrast and end-to-end success diverge** (§6). Under our one-step revision protocol, S2R raises the correction rate while the revised retry stays below the same checkpoint's fresh draw; LoRA supervised fine-tuning on retry prompts turns J positive on its fixed failure set in each of three seeds, but not detectably on the model's own errors, and, in all three seeds, lowers fresh-draw success below the untrained model's, as it does end-to-end success in the seed tested.

2 Separating the verdict from the attempt

Two retries, one verdict. A model, a problem and an answer y the verifier rejected, read as its symbolic equivalence class (the *answer class*), form a failed state. From each frozen failed state two retries run (Table 1). Both use the same solver, decoding and output-token ceiling under one controller that redraws any output in y's class. The controller rejects a completed answer in that class after it is generated, and bills it; it does not stop the solver from producing one. Both prompts hold the problem, an assistant turn and an instruction to find any error and produce an improved solution. With the *transcript kept*, the assistant turn is the whole failed transcript; with the *transcript withheld*, it is the line "[Previous solution withheld.]" and the instruction opens by saying that the earlier solution is not shown. The retention contrast J is the kept retry's success minus the withheld one's. Neither prompt says that the answer was wrong: the verdict acts through the controller alone. Because the withheld instruction differs by its opening clause, the comparison at one instruction and matched prompt length is the content ablation of §3.

Models and inference. The shared-controller study runs Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on 161 problems at an 8k ceiling, Gemma-3-12B-it (Gemma Team, 2025) on 85 at 8k, and Qwen3-8B and Qwen3-32B with thinking off (Yang et al., 2025a) on 100 each at 16k, on exact-answer mathematics screened per model (Appendix K). Each problem has two of the model's own wrong draws frozen as starts and eight episodes per start and condition; every solver draw, here and in the nine-policy collection (§5), is sampled at temperature .8 and top-p .95 and billed. Contrasts are paired by problem, with problem-level bootstrap 95% intervals. The post-training study (§6) reruns J from Llama-3.1-8B on its own sealed problem set, under the controller of Table 3 and with a note in both prompts that the answer is incorrect.

The value of the verdict alone.

Definition 1 (Failed state, reference law, verdict-only reference success)**.** Fix a model and a problem with correct answer $y^\star$. The *fresh-draw law* p_0 is the model's clean-context answer distribution on that problem and $q_0 := p_0(y^\star)$ its success rate. Let $y \neq y^\star$ be an answer the verifier has rejected, read as its answer class; the triple (model, problem, y) is a *failed state*. The *reference law* is p_0 conditioned off that class, $\rho := p_0(\cdot \mid \cdot \neq y)$, and its success rate, the *verdict-only reference success*, is

$$v_y \;:=\; \rho(y^\star) \;=\; \frac{q_0}{1 - p_0(y)}. \tag{1}$$

Lemma 1 (The verdict-only restart)**.** *If $p_0(y) < 1$, a restart B that resamples from p_0, with y's class rejected controller-side and never shown, returns in one accepted draw a draw from ρ: its success is v_y, above a fresh draw's by $v_y - q_0 = q_0\, p_0(y)/(1 - p_0(y))$.*

The restart B that attains v_y is prompted with the bare problem, so it is not the withheld retry: it gives the verdict's idealized value and runs as a policy in the nine-policy collection (§5), whose reference is the fresh draw A, while J compares the kept and withheld revision retries. Proofs, and the idealized framework that the other formal results belong to, are in Appendices C and D; symbols in Appendix B.

Table 1: **Conditions of §§2–4.** What each retry's prompt holds after the problem, what its controller does, and the key used in the tables and figures; y: the rejected answer's class. *Revise*: an instruction to review the solution shown, find any error and improve it (*hidden*: told that it is not shown; texts in Appendix E). Shared controller: at most 32 calls, a completed draw in y's class rejected and billed, a truncated call ending the episode as a failure. The fresh draw A, the verdict-only restart B and the post-training study's two retention retries, whose prompts also say that the answer is incorrect, follow the redraw rule of Table 3; D_{ret} and $D_{\varnothing}$ are not variants of that table's judged-prefix policy D.

Condition	Retry prompt holds	Controller	Key
Retention contrast J (Figure 1a)			
transcript kept	the whole failed transcript; revise	bans y; shared controller	D_{ret}
transcript withheld	"[Previous solution withheld.]"; revise, hidden	as transcript kept	$D_{\varnothing}$
Content ablation (Figure 1b, §§3–4): one instruction, matched prompt length			
placeholder	a content-free placeholder; revise	as transcript kept	–
placeholder, answer	the placeholder, then the attempt's final-answer line; revise	as transcript kept	–
redacted transcript	the failed transcript, final answer redacted; revise	as transcript kept	–
other problem's	another problem's redacted failed transcript; revise	as transcript kept	–

Conventions. Conditions, problem lists, seeds, estimators and decision rules were frozen in dated lab records before each collection (internal freezes, not third-party registrations); unmarked results are *registered* in that sense, and a *sealed* problem set was fixed before any policy ran on it. Registered is not one strength of evidence: on Llama-3.1-8B the retention contrast is a discovery, a pre-specified contrast of a transcript $\times$ hint design whose target was the interaction, with part of its data read on a 64-problem development set before the remaining 97 problems were collected; on Gemma-3-12B, Qwen3-8B and Qwen3-32B it is a prospective replication with J as the registered primary contrast (Appendix K). A dagger (†) marks a post hoc value; *offline* analyses reuse collected episodes and are named once per paragraph; each contrast's status is in Appendix F. An interval that contains zero is *not detected*. Conditions have plain names in the text and key letters in tables and figures (Tables 1 and 3).

3 KEEPING THE FAILED TRANSCRIPT LOWERS SUCCESS

Keeping the transcript lowers success in all four models, by .106 to .235, every interval excluding zero (Figure 1a): the withheld retry succeeds in .568, .575, .697 and .735 of episodes on Llama-3.1-8B, Gemma-3-12B, Qwen3-8B and Qwen3-32B, the kept one in .333, .451, .577 and .629 (Appendix K). J is also negative on the untrained base of the post-training study, under that study's controller and with both prompts saying that the answer is incorrect (Table 5).

Two simpler explanations predict the same sign: the kept prompt is longer, and it shows the rejected answer again. The content ablation tests both at one instruction and matched prompt length. Its 2×2 fills the assistant turn with either the failed transcript with its final answer redacted (every boxed expression, and the answer string in its last three non-empty lines) or a content-free placeholder, and shows or omits the rejected answer (Figure 1b; conditions in Table 19). A long content-free turn does not produce the loss: under the revision instruction the placeholder scores above the withheld retry on Llama-3.1-8B ($+.039$) and is not detected against it on the other three models (offline, specified after the results; Appendix K). The redacted transcript does: it stays below the placeholder in every model, by .074 to .260, and on the problems whose redacted transcripts nowhere contain the answer string it stays below in three models and is not detected on Qwen3-8B (Appendix K). Adding the answer back to it is not detected in any model; shown on the placeholder, the answer costs Llama-3.1-8B .177 and is not detected elsewhere.

Table 2: **Where keeping the transcript loses.** Each entry is a difference in the share of episodes with one outcome, transcript kept minus withheld ($D_{\text{ret}} - D_{\varnothing}$); problem-paired bootstrap 95% intervals; offline, specified after J was read; the J column carries the intervals of Appendix K. Success: correct at the first call (the retry with no rejection enforced) or at a later call (what the redraws add). Failure: *budget*, the ceiling reduced or spent by rejected calls (two registered states merged after the first run, Appendix G); a final call truncated at the full ceiling; an unparseable output; the 32-call cap. The blocks sum to J and $-J$, up to rounding. Gemma-3-12B at 16k and the code study: Table 21.

	success			failure				
						other terminal failures		
Collection	first call	later calls	total $= J$	accepted wrong	budget	full ceiling	unparseable	call cap
Llama-3.1-8B, 8k	$-.286$	$+.051$	$-.235$	$+.132$	$+.023$	$+.037$	$+.042$	.000
	$[-.329, -.245]$	$[+.038, +.066]$	$[-.276, -.194]$	$[+.095, +.168]$	$[+.015, +.032]$			
Gemma-3-12B, 8k	$-.307$	$+.182$	$-.124$	$-.043$	$+.165$	.000	.000	$+.002$
	$[-.378, -.246]$	$[+.145, +.225]$	$[-.179, -.071]$	$[-.107, +.021]$	$[+.114, +.221]$			
Qwen3-8B, 16k	$-.364$	$+.244$	$-.120$	$-.019$	$+.141$	$-.008$	.000	$+.006$
	$[-.418, -.308]$	$[+.189, +.294]$	$[-.175, -.065]$	$[-.062, +.026]$	$[+.096, +.189]$			
Qwen3-32B, 16k	$-.314$	$+.209$	$-.106$	$+.041$	$+.064$	$-.001$	.000	$+.002$
	$[-.367, -.266]$	$[+.158, +.261]$	$[-.157, -.058]$	$[-.001, +.081]$	$[+.024, +.105]$			

4 WHY THE LOSS SURVIVES THE BAN

The controller bans the rejected class, yet the kept transcript still costs success. Splitting J by call shows where (Table 2). At the first call, keeping the transcript costs .286 to .364 in all four models, and the transcript sends that call back to the rejected class far more often: the return rate rises by .173 on Llama-3.1-8B and by .391 to .497 on the other three (Appendix K). The redraws recover part of the cost, $+.051$ on Llama-3.1-8B and $+.182$ to $+.244$ on the others. What remains differs by model. On Gemma-3-12B and Qwen3-8B it sits in the budget states: each return is a billed call, and repeated returns spend the ceiling. On Qwen3-32B the budget states carry .064 of the .106. On Llama-3.1-8B more than half of the loss is an accepted answer that is new and wrong, which a ban on the old class cannot reach.

The return survives redaction. Either part of the transcript raises the first call's return in every model, and on Gemma-3-12B and both Qwen3 models the redacted transcript raises it more than the answer line (Table 19). Strict redaction leaves the answer string somewhere in 36.5 to 61.8% of the transcripts; on the problems where it occurs nowhere, the redacted transcript's first call still returns to the rejected class more than twice as often as the placeholder's, in all four models (post hoc on Llama-3.1-8B; Appendix K). Another problem's redacted transcript of matched length, in the same turn, leaves the first call's return not detectably above the placeholder's rate in any model; on Llama-3.1-8B, where the retry keeps to its own problem, the return is specific to its own transcript. Another problem's transcript costs success without the return: on Llama-3.1-8B about a fifth of the redacted transcript's loss, through the same route, a first call that leaves the class but is correct less often, and on the other three models the retry answers the donor's problem instead in .071 to .588 of episodes (post hoc; Table 20).

Routing and proposal. The two parts seen in Table 2, returning to the rejected class and leaving it for a wrong answer, split any retry policy's gap to a fresh draw exactly. For a retry policy φ that continues from a failed state and ends on one graded answer Z_φ (Definition 2), write $c_\varphi := \mathbb{P}(Z_\varphi = y^\star)$ for its success and $g_\varphi := \mathbb{P}(Z_\varphi \neq y)$ for its *leave rate*.

Proposition 1 (Routing/proposal ledger). *Let A be the fresh draw at the same ceiling, and on states where A leaves the class ($g_A > 0$) estimate v_y by c_A/g_A, which for a single clean-context draw is Eq. (1). For every policy φ on such a failed state, with routing and proposal terms $R_\varphi := (g_\varphi - g_A)\, v_y$ and $P_\varphi := c_\varphi - g_\varphi v_y$,*

$$c_\varphi - c_A \;=\; \underbrace{R_\varphi}_{\textit{routing}} \;+\; \underbrace{P_\varphi}_{\textit{proposal}} \tag{2}$$

exactly on every state, including $g_\varphi = 0$, where $R_\varphi = -g_A v_y$ and $P_\varphi = 0$; in particular $P_A = 0$.

Routing is the cost of not leaving the rejected class; proposal is the quality of the new answer when the retry does leave. For the idealized restart of Lemma 1, exclusion changes routing only: $P_B = 0$, so its gain over a fresh draw is all routing; under a finite budget the ban's costs are the budget states of Table 2. Under the shared controller, Llama-3.1-8B's first call that leaves the class is correct less often with the transcript kept (Appendix K): a negative proposal term, the part of its loss that the ban cannot recover. In the idealized framework, a retry call that can only copy y, redraw from the fresh-draw law or move toward answers other than $y^\star$ is correct at most at the rate v_y when it changes its answer, whereas a sampler whose beliefs are calibrated and updated on the verdict is correct at least at that rate on average (Theorem 1 and Proposition 2, Appendix C).

5 Complete retry policies

J isolates the transcript under one controller; a deployed retry policy changes instruction, exclusion and stopping along with what it keeps. A separate collection therefore prices complete retry policies that retain different parts of the attempt, each against a fresh draw at the same output-token ceiling on shared frozen wrong starts, under the controller of Table 3 rather than the shared controller of §§2–4. It runs Llama-3.1-8B, Gemma-3-12B and Qwen3-32B (Llama, Gemma and Qwen below; thinking off; Qwen runs the endpoints and checked facts) on exact-answer mathematics (MATH-500, Lightman et al., 2024; MATH train levels 2–5, Hendrycks et al., 2021; AIME 2024–2025), 80 sealed problems per family at 2k, 4k and 8k, two starts per problem, eight episodes per policy and start, and problem-level bootstrap intervals from 2,000 resamples. Retention runs from none to the whole answer history, an author-designed full-history continuation baseline inspired by the continuation stage of Bilal et al. (2026), not a reproduction of their refinement procedure (Appendix E). Each gap prices a whole policy, not the attempt's content, and none is pooled with J.

Across policies. The whole history is the worst policy in every family and ceiling, .33 to .48 below a fresh draw (Figure 2 and the first six rows of Table 4), while the verdict-only restart is not detected below the fresh draw in any setting where it ran. On Llama every content and history policy loses to the fresh draw, the stand-in least and the whole history most (Table 10; 24 of the 48 Llama/Gemma contrasts survive a family-wise adjustment, Appendix I); on Gemma and Qwen checked facts are not separated from the fresh draw. Proposition 1 splits these gaps state by state (exploratory and offline, specified after the aggregate contrasts; it describes behavior, not internal copying or belief, Proposition 4; Figure 3): most of the whole history's gap is routing, a share of .70 to 1.02 over the nine family–ceiling settings, and on Llama and Qwen its proposal term is also negative at every ceiling, so when it does leave the rejected class its new answer is correct less often than a fresh draw outside that class (Appendix H). At the first accepted call, where the whole history and the last output hold the same transcript and differ in instruction and attempt label, the whole history trails the last output by about .1 (offline; Appendix H). Checked equalities from the attempt, shown with the rest withheld, are not detected to help at the first call, before any redraw; most states yield no admissible equality, 127 of 160 on Llama, and the same slot filled with the problem's own reference-solution facts is not detected above an empty memo, so the null concerns this extraction pipeline (Appendix H).

Regimes. Three further inference regimes test the two endpoints on problems screened per regime from AIME 2022–2026, HMMT February 2025–2026 and MATH-500 (Appendix E): Gemma-4-31B-it (Gemma Team, 2026) with thinking on at 16k, and gpt-oss-120b (OpenAI, 2025) at low and at high reasoning effort at 32k; Qwen3-32B's thinking switch adds a fourth. The whole history stays below the fresh draw in every setting but one, gpt-oss-120b at low effort rerun on the high-effort problems, which has 65.1% of the whole history's episodes at the 64-call cap and is negative with cap

Table 3: **The nine-policy collection** (§5), in the columns of Table 1; nine retry policies priced against the fresh draw A. *Continue*: carry the solution shown forward without starting over; *told wrong*: a note says that the previous answer is incorrect. Controller: a banned or truncated draw is redrawn, the seventh consecutive draw accepted, at most 64 calls per episode. *Judge, extractor and checker tokens are not billed; every solver draw is.

Condition	Retry prompt holds	Controller	Key
Resampling and stopping controls			
fresh draw	the problem only	no ban; first complete draw	A
verdict-only restart	the problem only	bans y; first accepted draw	B
fresh draws to the ceiling	the problem only	no ban; runs to the ceiling	G
restart to the ceiling	the problem only	bans y; runs to the ceiling	B_{budget}
Content handling			
content-free stand-in	a stand-in of the attempt's length; revise, hidden	bans y; runs to the ceiling	E
checked facts*	first call: the stand-in and ≤ 3 checked equalities from the attempt, told wrong; revise, hidden; then as last output	no ban; runs to the ceiling	F
judged prefix*	the prefix of the attempt a judge locates as correct; no instruction	no ban; runs to the ceiling	D
Retaining history			
last output	the previous output; revise	no ban; runs to the ceiling	C_{last}
transcript + facts*	first call: the failed transcript and ≤ 3 checked equalities, told wrong; revise; then as last output	no ban; runs to the ceiling	H
whole history	every earlier output, as "[Attempt 1]", "[Attempt 2]", . . . ; continue	no ban; runs to the ceiling	C_{full}

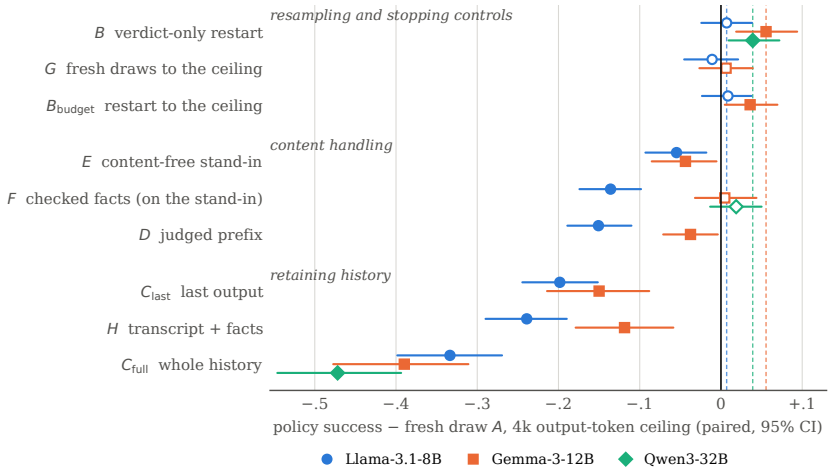

Figure 2: **Retry policies at a shared 4k output-token ceiling.** Policy success minus the fresh draw A; each label gives the policy's key letter and plain name (Table 3); each gap prices a whole policy, instruction, ban and stopping included; paired problem-level bootstrap 95% intervals; filled marker: interval excludes zero. Dashed vertical lines: each family's verdict-only restart minus the fresh draw, $B - A$. All three ceilings are in Appendix H.

hits scored as failures. Within one model, at 16k on one problem set, thinking raises Qwen3-32B's fresh draw from .69 to .92 and moves the whole history's gap from $-.458$ to $-.126$ (descriptive; Appendix G); across reasoning effort the direction depends on the problem set (Table 14).

Table 4: **The two endpoints across budgets and inference regimes.** Policy minus the fresh draw A at the same ceiling, paired problem-level bootstrap 95% intervals; B: verdict-only restart; C_{full}: whole history; last column: cap hits scored as failures, where that moves the contrast by more than .01. Qwen3-32B's 2k and 8k rows extend its single-ceiling registration; truncation rates in Appendix G; –: not run.

Regime, ceiling	$B - A$	$C_{\text{full}} - A$	$C_{\text{full}} - A$, cap hit = failure
Llama-3.1-8B, 2k	$+.035$ $[+.009, +.061]$	$-.334$ $[-.395, -.272]$	
Llama-3.1-8B, 8k	$+.024$ $[-.011, +.058]$	$-.341$ $[-.398, -.286]$	
Gemma-3-12B, 2k	$+.039$ $[+.008, +.070]$	$-.395$ $[-.473, -.309]$	
Gemma-3-12B, 8k	$+.073$ $[+.039, +.109]$	$-.338$ $[-.429, -.247]$	
Qwen3-32B, 2k	–	$-.459$ $[-.534, -.383]$	
Qwen3-32B, 8k	–	$-.479$ $[-.554, -.400]$	$-.512$ $[-.585, -.438]$
Qwen3-32B thinking, 16k	$-.005$ $[-.020, +.012]$	$-.126$ $[-.184, -.070]$	
Gemma-4-31B thinking, 16k	$+.015$ $[-.018, +.050]$	$-.147$ $[-.219, -.075]$	
gpt-oss-120b low, 32k	$+.084$ $[+.054, +.116]$	$-.300$ $[-.370, -.229]$	$-.457$ $[-.532, -.380]$
gpt-oss-120b high, 32k	$+.016$ $[-.017, +.048]$	$-.101$ $[-.175, -.032]$	
gpt-oss-120b low, high-effort problems, 32k	$-.006$ $[-.036, +.025]$	$+.091$ $[+.004, +.185]$	$-.158$ $[-.234, -.084]$

Table 5: **Post-training, sealed test.** 256 problems $\times$ 8 episodes. A: fresh draw; $D_{\varnothing}$, D_{ret}: transcript withheld, kept; Δ: change against the untrained base. LS, LR, FS: LoRA supervised fine-tuning, LoRA GRPO and full-parameter supervised fine-tuning on retry prompts. 99% Bonferroni intervals on J for the trained models, 95% elsewhere; first LS seed (two further seeds in Appendix L).

Model	A	$D_{\varnothing}$	D_{ret}	$J = D_{\text{ret}} - D_{\varnothing}$	ΔA	$\Delta D_{\varnothing}$	ΔD_{ret}
base	.522	.486	.261	$-.226$ $[-.262, -.189]$	–	–	–
LS	.354	.374	.418	$+.044$ $[+.010, +.079]$	$-.169$ $[-.203, -.135]$	$-.112$ $[-.148, -.075]$	$+.157$ $[+.117, +.196]$
LR	.542	.494	.282	$-.211$ $[-.260, -.164]$	$+.020$ $[-.003, +.042]$	$+.007$ $[-.013, +.029]$	$+.021$ $[-.002, +.045]$
FS	.180	.196	.268	$+.072$ $[+.044, +.100]$	$-.343$ $[-.381, -.303]$	$-.290$ $[-.328, -.251]$	$+.007$ $[-.030, +.044]$

6 Do the metrics agree after training?

They do not. Here the *correction rate* is the success of a retry that sees its failed attempt and is told that the answer is incorrect.

Self-correction RL. S2R's process- and outcome-level RL checkpoints and their Llama-3.1-8B base were retried from the same frozen wrong starts of 80 problems under our one-step revision protocol, one revision call told that the answer is incorrect (S2R's own evaluation is greedy verify-then-correct; Appendix J). Training raises revision success from .209 to .475 and .480, but a fresh draw from each checkpoint's clean context rises too, to .537 and .543, and revision stays below it at

Table 6: **End to end, base against LS** (first seed; from Table 24). Retry: one episode from a complete wrong first attempt, conditioned on each model's own errors, so the populations differ. End to end: the first attempt is correct, or wrong and then corrected by the retry.

	first attempt	retry success			end to end		
Model	correct	A	$D_{\varnothing}$	D_{ret}	$\mathcal{E}_A$	$\mathcal{E}_{D_{\varnothing}}$	$\mathcal{E}_{D_{\text{ret}}}$
base	.650	.340	.323	.194	.739	.735	.701
LS	.460	.292	.316	.314	.604	.616	.615

every checkpoint (Figure 4 and Table 16, Appendix J). With a 16k per-call cap, where the outcome-RL revision's truncation falls from .205 to .011 against a same-node 4k control, it is still below the collected fresh draw ($-.062$).

Retry fine-tuning. The same base was fine-tuned on retry prompts holding its own wrong starts under three recipes, low-rank adaptation (LoRA; Hu et al., 2022) supervised fine-tuning (LS), full-parameter supervised fine-tuning (FS) and LoRA group relative policy optimization (GRPO; Shao et al., 2024; LR). Each checkpoint was chosen among exports at 25, 50 and 100% of training by validation success with the transcript kept. The fresh draw and the two retention retries were then run on a sealed test set of the base's wrong starts, with the result rows written into a frozen analysis script before the training data existed (Table 5). Both supervised recipes turn J positive on this fixed failure set, and from the two sides differently: after LS the kept retry rises and the withheld one falls, after FS the kept retry stays where the base left it and only the withheld one falls (Table 5; two further LS seeds in Table 22). The same recipes lower the model's own fresh-draw success, also with direct-solution rows mixed in and at fp32 as at bf16 serving (Appendix L); no tested policy of either supervised model reaches the base's fresh draw. J cannot be read alone: training selected on the kept retry raised J while it lowered the level of success, so J by itself would report progress for models that solve fewer problems. LR's adapter barely left the base (per-token Kullback–Leibler estimate about 4×10^{-4}) and is not a test of GRPO. The positive J is established on the fixed failure set only: on each model's own wrong starts it is not detected (Table 23). End to end, on 400 problems no failure filter has touched, the base solves more with a fresh draw than LS does under any of its three policies, LS's kept-minus-withheld difference is not detected, and most of the gap is already in the first attempt (Table 6; first seed).

7 Implications and boundaries

Enforcing the rejection and keeping the failed attempt are separate decisions. Enforcement is cheap and can be a controller default: banning y's class gives the answer law closest to the model's own, in $\mathrm{KL}(q \,\|\, p_0)$, among those that put no mass on y (Proposition 3, for the idealized controller), and adds $+50/+54/+72$ billed tokens per state over the fresh draw on Llama and $+68/+198/+329$ on Gemma across the three ceilings (offline, from the episode table of Appendix E). End to end, at base, the withheld retry, revision prompt included, is not separated from the fresh draw ($-.004$; Appendix L). Retention is the expensive decision: under the shared controller the kept retry spends about twice the withheld one's output tokens per episode, 3.7k to 5.8k against 1.8k to 2.5k on the three further models (Appendix K), and a policy that runs to the ceiling bills the whole ceiling (Appendix H). Whether the attempt enters the next context should rest on its retention contrast, read with the level of success it comes with, not on the computation already spent.

Boundaries. The claims are for exact-answer mathematics under a verifier whose rejections are correct, and for released models not trained for retry apart from the checkpoints of §6. On code the retention contrast is not detected ($-.017$; $-.042^{\dagger}$ once a parse rule that was not policy-neutral is relaxed; Appendix K). When a verifier can reject a correct answer, the coherent update down-weights the rejected class instead of banning it (Proposition 5), a case the numbers here do not price; our rejections are correct up to answer normalization (Appendix K). Family, size and generation are confounded, and the four shared-controller models differ in problem set and ceiling; the post-training study is one base and one dataset, trained on kept-transcript prompts, with three seeds for LS only.

Conclusion. With the rejection enforced on both sides and every solver draw billed, what the failed attempt adds beyond the verdict was negative in the four models tested, also with its final answer redacted, and training that reversed that sign did not raise the level of success. The rejection tells the next attempt where not to go: correcting needs that verdict, and reusing the failed attempt has to earn its place beyond it.

AI USE STATEMENT

Generative AI tools were used under the authors' supervision, with author control before, during and after each use: for research ideation or execution, to work out and check ideas and proofs the authors originated and steered, to provide feedback on the experiments, as coding assistance, and to verify the experimental records and documentation; for retrieval and discovery, to search and summarize related work and to check citations; to draft sections of the paper from author-specified content and results; and to aid or polish the writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates and fine-tunes open-weight language models on public mathematics and code benchmarks; it involves no human subjects, no personal data, and no new data collected from people. The deployment risk the findings bear on is treating an imperfect verifier as an oracle: every result here assumes rejections true up to answer normalization, and §7 states what changes when a rejection can be false.

REPRODUCIBILITY STATEMENT

Appendix E gives the models, problem sets, decoding, grading and controller rules and the provenance of every number; Appendix F defines each contrast and its estimator; Appendix G gives the registrations and every deviation from them; Appendices H–L report every condition; and Appendices C and D give the full statements and proofs of the supporting theory. The analysis, table and figure scripts with the stored analysis outputs every reported number is read from, the prompt templates, problem lists and start sets, the grader, and the weights of the first-seed post-training checkpoints and of the mixed-data model of Appendix L will be released with the paper.

REFERENCES

Johin Johny Arimbur. How many tries does it take? Iterative self-repair in LLM code generation across model scales and benchmarks. *arXiv preprint arXiv:2604.10508*, 2026.

Shelly Bensal, Umar Jamil, Christopher Bryant, Melisa Russak, Kiran Kamble, Dmytro Mozolevskyi, Muayad Ali, and Waseem AlShikh. Reflect, retry, reward: Self-improving LLMs via reinforcement learning. *arXiv preprint arXiv:2505.24726*, 2025.

Ahsan Bilal, Muhammad Ahmed Mohsin, Muhammad Umer, Lena Trigg, Ali Subhan, Muhammad Ali, and Dean F. Hougen. Refining over resampling: Test-time self-correction for LLM reasoning. *arXiv preprint arXiv:2608.05643*, 2026.

David Blackwell. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Isadore Blumen, Marvin Kogan, and Philip J. McCarthy. *The Industrial Mobility of Labor as a Probability Process*. Cornell University Press, 1955.

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.

Jiefeng Chen, Jie Ren, Xinyun Chen, Chengrun Yang, Ruoxi Sun, Jinsung Yoon, and Sercan Ö. Arık. SETS: Leveraging self-verification and self-correction for improved test-time scaling. *Transactions on Machine Learning Research*, 2025. arXiv:2501.19306.

Kuan-Yen Chen, Fang-Yi Su, Shih-Yen Lin, Bao Li, and Jung-Hsien Chiang. The self-correction illusion: Role relabeling gates explicit error flagging in large language models. *arXiv preprint arXiv:2606.05976*, 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.

Sijia Chen and Baochun Li. Toward adaptive reasoning in large language models with thought rollback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2412.19707.

Teng-Ruei Chen. Resample or reroute? Budget-aware test-time model selection for large language models. *arXiv preprint arXiv:2607.08665v2*, 2026. v3 is retitled "Recoverable Stopping Debt Without Identified Action Selection".

Yun Cheng, Xingyu Zhu, Haoyu Zhao, and Sanjeev Arora. Contextual drag: How errors in the context affect LLM reasoning. *arXiv preprint arXiv:2602.04288*, 2026.

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

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

Bartłomiej Cupiał, Jan Łojek, Mikołaj Garstecki, Szymon Pobłocki, Alicja Ziarko, and Piotr Miłoś. What drives interactive improvement from feedback? In *RLxF Workshop, International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.30774.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1-2):12–22, 1983.

Yifeng Ding and Lingming Zhang. SWE-Replay: Efficient test-time scaling for software engineering agents. *arXiv preprint arXiv:2601.22129*, 2026.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Leo A. Goodman. Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56(296):841–868, 1961.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.11738.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Soham De, Larisa Markeeva, Petar Veličković, and Razvan Pascanu. Understanding performance gap between parallel and sequential sampling in large reasoning models. *arXiv preprint arXiv:2604.05868*, 2026.

Esmail Gumaan. Feedback that backfires: Why small language model agents repeat the call they just watched fail. *arXiv preprint arXiv:2608.23651*, 2026.

Bradley Guo, Jingwen Gu, Jin Peng Zhou, and Wen Sun. Learning to self-correct through chain-of-thought verification. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! (PUT) at ICML 2025*, 2025. URL https://openreview.net/forum?id=AbO4lCvlo3.

Kim Hammar, Tansu Alpcan, and Emil C. Lupu. Optimal stopping of self-refining foundation models. In *65th IEEE Conference on Decision and Control (CDC)*, 2026. arXiv:2608.10729.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. GLoRe: When, where, and how to improve LLM reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jie Hu. Oracle gap and signal fidelity: A fixed-pool diagnostic for test-time collaboration. *arXiv preprint arXiv:2607.17531*, 2026.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *International Conference on Learning Representations*, 2025. arXiv:2412.01951.

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 (ICLR)*, 2024.

Mehmet İşcan. Falsification, not exposure: An internally preregistered placebo-controlled decomposition of self-repair feedback in frozen small code models. *arXiv preprint arXiv:2606.31511*, 2026.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination-free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Dongwei Jiang, Alvin Zhang, Andrew Wang, Nicholas Andrews, and Daniel Khashabi. Feedback friction: LLMs struggle to fully incorporate external feedback. *arXiv preprint arXiv:2506.11930*, 2025a.

Dongwei Jiang, Jingyu Zhang, Orion Weller, Nathaniel Weir, Benjamin Van Durme, and Daniel Khashabi. SELF-[IN]CORRECT: LLMs struggle with discriminating self-generated responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24266–24275, 2025b. arXiv:2404.04298.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. arXiv:2406.01297.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.12917.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph 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 (SOSP)*, 2023.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment. *arXiv preprint arXiv:2311.08596*, 2023.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Jipeng Li, Zeyu Gao, Yubin Qi, Hande Dong, Weijian Chen, and Qiang Lin. Unleashing the true potential of LLMs: A feedback-triggered self-correction with long-term multipath decoding. *arXiv preprint arXiv:2509.07676*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.91. URL https://aclanthology.org/2024.findings-acl.91/. arXiv:2402.14809.

Licheng Liu, Zihan Wang, Linjie Li, Chenwei Xu, Yiping Lu, Han Liu, Avirup Sil, and Manling Li. A simple "try again" can elicit multi-turn LLM reasoning. *arXiv preprint arXiv:2507.14295*, 2025.

Yuanxin Liu, Ruida Zhou, Xinyan Zhao, Amr Sharaf, Hongzhou Lin, Arijit Biswas, Mohammad Ghavamzadeh, Zhaoran Wang, and Mingyi Hong. REVES: REvision and VErification–Augmented training for test-time scaling. *arXiv preprint arXiv:2606.18910*, 2026.

Huanhuan Ma, Henry Peng Zou, Chengze Li, Enze Ma, Yunyue Su, and Philip S. Yu. Sycophancy suppression can impair rational updating: Anti-sycophancy should preserve the ability to update. *arXiv preprint arXiv:2608.26511*, 2026. Findings of EMNLP 2026.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S^2R: Teaching LLMs to self-verify and self-correct via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22632–22654, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.17651.

Lovish Madaan, Aniket Didolkar, Suchin Gururangan, John Quan, Ruan Silva, Ruslan Salakhutdinov, Manzil Zaheer, Sanjeev Arora, and Anirudh Goyal. Rethinking thinking tokens: LLMs as improvement operators. *arXiv preprint arXiv:2510.01123*, 2025.

Parsa Mazaheri and Kasra Mazaheri. Prior audit-repair context shifts LLM verifier thresholds toward leniency. *arXiv preprint arXiv:2608.16003*, 2026.

Iliya Mirzaei. Sample more, reflect less: Self-refine and Reflexion lose to repeated sampling at equal token cost, from 1.5B to 7B. *arXiv preprint arXiv:2607.28576*, 2026.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Prashanti Nilayam, Kiran Ramanna, and Prashil Tumbade. Detection without correction: A two-parameter decomposition of multi-stage LLM pipelines. *arXiv preprint arXiv:2605.27559*, 2026.

Jingjie Ning, Xueqi Li, and Chengyu Yu. Revision or re-solving? Decomposing second-pass gains in multi-LLM pipelines. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026. arXiv:2604.01029.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.09896.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. doi: 10.1162/tacl_a_00660. URL https://aclanthology.org/2024.tacl-1.27/. arXiv:2308.03188.

Nearchos Potamitis and Akhil Arora. Are retrials all you need? Enhancing large language model reasoning without verbalized feedback. *arXiv preprint arXiv:2504.12951*, 2025.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.18219.

Mohammadreza Saadat and Steve Nemzer. Certainty robustness: Evaluating LLM stability under self-challenging prompts. *arXiv preprint arXiv:2603.03330*, 2026.

Ankur Samanta, Akshayaa Magesh, Ayush Jain, Kavosh Asadi, Youliang Yu, Daniel Jiang, Boris Vidolov, Kaveh Hassani, Paul Sajda, Jalaj Bhandari, and Yonathan Efroni. Structure enables effective self-localization of errors in LLMs. *arXiv preprint arXiv:2602.02416*, 2026.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aman Sharma and Paras Chopra. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute. *arXiv preprint arXiv:2511.02309*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.13548.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.11366.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.03314.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*, 2023. Presented at the NeurIPS 2023 Workshop on Foundation Models for Decision Making.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*, 2024.

Hao Tang, Keya Hu, Jin Peng Zhou, Sicheng Zhong, Wei-Long Zheng, Xujie Si, and Kevin Ellis. Code repair with LLMs gives an exploration-exploitation tradeoff. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.17503.

Yefan Tao, Gerald Friedland, Madhusudhanan Chandrasekaran, and Luyang Kong. Reflection or re-generation? Why LLM revision fails where human revision succeeds. *arXiv preprint arXiv:2607.28908*, 2026.

Ken Tsui. Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models. *arXiv preprint arXiv:2507.02778*, 2025. Accepted at COLM 2026.

Yuvraj Verma. Try again, don't look back: Blind resampling outperforms self-repair in small code models. *arXiv preprint arXiv:2607.26117*, 2026.

Chenyu Wang, Yunbo Lyu, Junda He, Zhou Yang, Chenxing Zhong, Yaniv Harel, and David Lo. Fail-fast, restart-smart: Early failure prediction and restart for SWE agentic tasks. *arXiv preprint arXiv:2608.03222*, 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*, pp. 19916–19939, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1112. URL https://aclanthology.org/2024.emnlp-main.1112/. arXiv:2406.06461.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2211.00053.

Yotam Wolf, Noam Wies, and Amnon Shashua. When does in-context search help? A sampling-complexity theory of reflection-driven reasoning. *arXiv preprint arXiv:2607.06720*, 2026.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=VNckp7JEHn. arXiv:2408.00724.

Yitao Wu, Si Shen, Rui Yang, Hong Peng, and Bin Hu. Verify, repair, repeat, or stop? Robust stopping for noisy verify-repair loops in LLM agents. *arXiv preprint arXiv:2607.17641*, 2026.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Enhancing mathematical reasoning in LLMs by stepwise correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 21602–21623, 2025b. arXiv:2410.12934.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models' vacillations in judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10709–10745, 2024. arXiv:2310.02174.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. *arXiv preprint arXiv:2402.11436*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Zhanfu Yang. Why retrying fails: Context contamination in LLM agent pipelines. *arXiv preprint arXiv:2605.08563*, 2026.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. Confidence v.s. critique: A decomposition of self-correction capability for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3998–4014, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.203. URL https://aclanthology.org/2025.acl-long.203/. arXiv:2412.19513.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.2, how to learn from mistakes on grade-school math problems. *arXiv preprint arXiv:2408.16293*, 2024.

Shen Yin, David Ken, and Joel Stremmel. Denoising iterative self-correction: Structured verification loops for reliable LLM reasoning. *arXiv preprint arXiv:2606.21724*, 2026.

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Yan Liu, and Han Qiu. Understanding the dark side of LLMs' intrinsic self-correction. *arXiv preprint arXiv:2412.14959*, 2024.

A Extended related work

Comparators built from fresh samples. Comparison against fresh draws is established practice. Saunders et al. (2022) compare critique-conditioned refinement with a best-of-8 baseline generated from scratch, which they call roughly compute-equalized, and find refinement only competitive with it at scale; Wang et al. (2024) charge reasoning strategies in tokens and find self-consistency at equal budget usually ahead, with Reflexion degrading as budget grows; Tang et al. (2024) frame code repair as an arm-acquiring bandit over which program to refine next, in which refining the empty program is a fresh draw. İşcan (2026) decomposes self-repair feedback in 0.5–1.5B code models on units no cached draw had solved, with placebo conditions matched in template and bullet count and a facts-only condition that withholds the failed code, alongside the instruction-only restart of §1. Verma (2026) runs a placebo condition that shows the failed program with a content-free notice, and finds that two solved exemplars from other tasks leave success unchanged (within ± 3.5 points), the nearest antecedent of the other-problem transcript control (Appendix K); Arimbur (2026) reports iterative self-repair gains on execution feedback concentrated in the first two rounds and greedy self-repair matching or beating temperature-0.8 resampling with fewer tokens in three of four models, Ding & Zhang (2026) recycle archived trajectories in software-engineering agents at lower cost than sampling from scratch, and Wang et al. (2026) restart a predicted failure without its prompt history while leaving the interrupted changes available, so the code-side record is mixed; Potamitis & Arora (2025) find plain retrial on a verifier's rejection competitive with structured reasoning frameworks. Here the verdict-only restarts are held to the rejection by the controller. At matched token cost, Mirzaei (2026) finds self-refinement never reliably beating repeated sampling up to 7B while Sharma & Chopra (2025) find chains conditioned on their own prior attempts beating parallel self-consistency in most matched-token configurations, both voting over completed chains with no verifier bit, and Chen et al. (2025) report better matched-token scaling for sampling with self-verification and self-correction, mainly on frontier API models; Bilal et al. (2026) refine each sampled rollout by self-critique before majority voting, with no verifier, and report gains over majority voting and verifier-based best-of-N at a fixed number of primary samples rather than matched compute. Tao et al. (2026) score oracle-signalled revision on multiple-choice questions against a random reshuffle among the remaining options; the verdict-only reference replaces that reshuffle with the model's own law off the rejected class. Wolf et al. (2026) treat the base model as a prior and reflection as a posterior update, and derive when sequential attempts beat parallel sampling as a sampling-complexity rate.

What to keep after a failure. Work on retention varies how much of a failure a retry keeps, or measures what a kept failure does to it. Beyond the pull toward similar errors (§1), Cheng et al. (2026) find that neither external feedback nor self-verification removes it and that their Filter condition does not recover the clean-context level in most settings; that condition is the nearest antecedent of policy F, whose kept facts are checked independently. The dominance result of Yang (2026) (§1) comes with a SWE-bench fit in which an i.i.d. model overstates the pass@3 of same-context retries. Prefix continuation is the policy of Samanta et al. (2026), who locate the first erroneous thought and resample from the prefix, and of Wu et al. (2025b), who revise from a verifier-located step, with the rest of the failed path still shown, and beat best-of-10 while spending 77.8% fewer tokens; Chen & Li (2024) roll back to an earlier thought and regenerate from the prefix with accumulated error analyses of the failed chain in the prompt, at about $35\times$ zero-shot output tokens on MATH. Policy D runs prefix continuation at equal budget. Ning et al. (2026) separate a cross-model second pass into re-solving, scaffold and content terms using a semantically null placeholder draft, and report that in code generation a null draft still scaffolds while weak draft content can be harmful (Appendix K gives this paper's code collection). Chen et al. (2026a) find that re-presenting a byte-identical copy of an erroneous claim from the model's own thought under an external role raises explicit correction by 23 to 93 points; in the shared-controller study the transcript and the placeholder sit in the same assistant turn. Madaan et al. (2025) refine from a bounded per-round workspace distilled from parallel drafts (a summary or a few selected drafts) instead of from all prior drafts, and Bulut (2026) finds, with restart controls, that continuing an existing reasoning path compresses compute more than it extends what is solved. Li et al. (2025) and Samanta et al. (2026) also spend a negative bit without showing the rejected answer, and Jiang et al. (2025a) reject previously attempted answers explicitly inside a retained-context feedback loop; here the comparator is a verdict-only restart held to the same rejection. Gumaan (2026) attributes most of a small agent's repetition of a failed tool call, persistence $1 - g_\varphi$ in the notation here, to the surface form of the recorded failure: describing

the failure instead of replaying the call removes most of the repetition, a decoder-level ban on failed strings acts on the same quantity, and deleting the failure to retry from the clean context repeats most. Here, redacting the rejected answer from a kept transcript does not stop the return (§4). Related decompositions split an unsignalled always-revise accuracy shift into content and parseability margins (Chen et al., 2026b) and a fixed-pool selector's net gain into recoverable mass, signal coverage, selection quality and harm to already-correct outputs (Hu, 2026), a selection-side counterpart of the routing/proposal ledger.

Measuring the conditional statistic. Nilayam et al. (2026) measure $\mathbb{P}(\text{correct} \mid \text{changed})$ across four model families and introduce a detection $\times$ correction two-factor frame; the reference they score it against is a fixed constant. Yang et al. (2025b) decompose self-correction into a confidence rate on previously correct answers and a critique rate on previously wrong ones, and in normalizing their relative score take first-attempt accuracy as the critique rate's ceiling; the verdict-only reference success exceeds first-attempt accuracy whenever $q_0 > 0$ and $p_0(y) > 0$ (Lemma 1). Lin et al. (2024) benchmark critique-and-correct with an oracle condition that corrects only responses labeled incorrect and find those corrections still short of direct generation for the stronger models. Jiang et al. (2025b) score a model's pick among its own generations against a random draw from the same pool and report that the pick does not beat the draw. The survey of Pan et al. (2024) lists the absence of a quantitative metric and of a theoretical account among the field's open problems, and that of Kamoi et al. (2024) finds research questions often left undefined and tested with ill-suited experiments. Tsui (2025) establishes the self- versus external-error blind spot on injected errors. Stratifying a challenge by whether the challenged answer was right has precedent in accuracy falling when a correct answer is challenged, under follow-up pressure, bare pushback or a fixed "you are wrong" prompt, and in refinement applied regardless of correctness (Sharma et al., 2024; Laban et al., 2023; Xie et al., 2024; Zhang et al., 2024; Saadat & Nemzer, 2026; Ma et al., 2026); Yin et al. (2026) gate corrections so that noisy verification does not rewrite already-correct answers. These results bear on the boundary of §7.

Stopping. Huang et al. (2024) attribute earlier intrinsic self-correction gains to oracle-label stopping, and Stechly et al. (2023; 2024) find that self-verification without an external checker does not help and can sharply degrade accuracy; Wu et al. (2026) show that externally imposed stopping policies for noisy verify–repair loops yield engineering gains, and Hammar et al. (2026) give the optimal-stopping formulation on a scalar quality score. Crossing exclusion with stopping in §5 (A, G, B, B_{budget}) separates the two factors those results bundle.

Self-correction and training. Self-correction loops report gains from iterating on the model's own output (Madaan et al., 2023; Shinn et al., 2023), and critic-guided loops (Gou et al., 2024) and verifier reranking (Cobbe et al., 2021; Lightman et al., 2024) add an external signal. On the training side, Welleck et al. (2023) train separate correctors, Kumar et al. (2025) train multi-turn self-correction with RL and Qu et al. (2024) with iterative reward-weighted fine-tuning, Ma et al. (2025) (the recipe measured in §6 and Appendix J) trains self-verify-and-revise behavior, and Bensal et al. (2025) reward a self-reflection written after a failure when the retry that sees it succeeds, with binary feedback only. Three trained-correction results point the other way from the reading here: Qu et al. (2024) compare five turns of trained revision with pass@5 at an equal number of oracle-verifier queries and report a sequential advantage; Kumar et al. (2025) report a batch of samples plus one correction round beating twice as many samples at a fixed budget under majority voting, with no verifier; and Liu et al. (2025), training multi-turn RL on a content-free "try again" bit with a penalty on repeated answers, find the multi-turn model up to 14% above the single-turn checkpoint's parallel sampling at comparable inference cost while keeping single-turn accuracy. Liu et al. (2025) also report that single-turn-trained models give the same answer across five turns in 70% of failed cases, the routing term priced here; Havrilla et al. (2024) find refinement alone barely helps and an external reranker carries the gain; and sequential retries beat parallel sampling only with a near-perfect hidden-state error detector (Ye et al., 2024) or a verification channel far richer than one bit (Guo et al., 2025). That one score can hide separate quantities (§6) has precedents: Liu et al. (2026) prove that policies with equal single-shot success can differ in sequential-revision success, and Cupiał et al. (2026), netting feedback gains against unguided self-refinement over thirteen open-weight models, find first-turn accuracy and feedback adherence to be different capabilities.

Test-time scaling. Repeated sampling raises coverage (Brown et al., 2024), and compute-optimal allocation of sampling, search and revision raises accuracy after selection (Snell et al., 2025; Wu et al., 2025a); sharpening-style selection (Huang et al., 2025) and self-consistency (Wang et al.,

2023) act on the decision layer, and budget-aware routing (Chen, 2026) chooses between resampling one model and rerouting to another, not between revising and resampling. Gu et al. (2026) test three accounts of why sequential sampling trails parallel sampling in reasoning models (the aggregator, the longer context, and lost exploration from conditioning on earlier answers) and find the third the main one; the terminal-state split of Appendix K prices that conditioning under a billed controller. Laban et al. (2025) measure the multi-turn form of the same reliance on an early attempt, and Xu et al. (2024) a self-bias that grows over refinement rounds. Mazaheri & Mazaheri (2026) find leniency in LLM verifiers holding an earlier audit-and-repair episode; the verdicts here come from an exact-answer or test-signature grader that never reads the transcript.

B NOTATION

Table 7: Symbols used in the main text and the appendices.

$\mathcal{Y}$, $y^\star$	finite answer space; unique correct answer
p_0, q_0	clean-context (fresh-draw) law; its success rate $q_0 = p_0(y^\star)$
y	the rejected answer, read as its symbolic equivalence class
ρ	reference law $p_0(\cdot)\mathbf{1}\{\cdot \neq y\}/(1 - p_0(y))$ (Definition 1)
v_y	verdict-only reference success, $q_0/(1 - p_0(y)) = \rho(y^\star)$; the verdict's increment is $v_y - q_0$
φ	a retry policy; the policies of Table 3 are A, B, G, B_{budget}, E, F, D, C_{last}, H, C_{full}, with their plain names; D_{ret}, $D_\varnothing$ of Table 1
A, B	the fresh draw; the verdict-only restart, which never sees the attempt and whose controller rejects y's class, prompted with the bare problem (Lemma 1, Table 3)
Z_φ	the graded answer of an episode of φ, its last completed call
c_φ, g_φ	success $\mathbb{P}(Z_\varphi = y^\star)$; leave rate $\mathbb{P}(Z_\varphi \neq y)$; $1 - g_\varphi$ is persistence
ξ_φ	changed-proposal excess $c_\varphi/g_\varphi - v_y$ (Definition 2)
R_φ, P_φ	routing $(g_\varphi - g_A)v_y$ and proposal $c_\varphi - g_\varphi v_y$ terms (Proposition 1)
Δc	the gap $c_\varphi - c_A$ that $R_\varphi + P_\varphi$ splits (Figure 3); in Table 5, Δ is a change against the base model
M_y	frontier margin $c_\varphi - v_y = g_\varphi \xi_\varphi - (1 - g_\varphi)v_y$ (Appendix J)
K, $\tilde{K}$	one-call kernel from y; its changed-conditional law $K(z\|y)/(1 - K(y\|y))$
λ, $\tilde{K}$	copy mass and work component in $K = \lambda\mathrm{Id} + (1 - \lambda)\tilde{K}$ (not identifiable; Lemma 2)
α_y, μ_y, m	anchoring strength; anchoring law; its exact-repeat share $\mu_y(y)$
$N(y)$, $N[y]$	a neighbourhood of y not containing $y^\star$; $\{y\} \cup N(y)$
Θ, X, Y, W	random truth; context; the rejected answer as a random draw $Y \sim p_0$; verifier bit $\mathbf{1}\{Y = \Theta\}$
$\mathcal{B}$, $\mathcal{G}$	belief-consistent sampler class; its information σ-algebra
I_0, β, Π	$\sigma(X, Y, W)$; the belief $\mathbb{P}(\Theta = \cdot \mid \mathcal{G})$ and the action law of Appendix D
$\mathcal{H}$, $\rho_{\mathcal{H}}$	a refuted set containing y; p_0 conditioned off it (Proposition 3)
TPR, TNR	a noisy verifier's rates (Proposition 5)
$\bar{\xi}_{\text{info}}$, $\bar{\xi}_{\text{pol}}$, $\bar{\xi}$	information and policy components of the population changed-proposal excess (Theorem 5)
b	output-token ceiling of an episode, $b \in \{2{,}048,\ 4{,}096,\ 8{,}192,\ 16{,}384,\ 32{,}768\}$
D_{ret}, $D_\varnothing$	the whole failed transcript kept, or withheld behind the one-line withholding notice "[Previous solution withheld.]" (the *withheld* condition, with the hidden revision instruction), under one exclusion controller, in the shared-controller study and the post-training study (§2; D_{ret} is not the judged prefix D)
J	$D_{\text{ret}} - D_\varnothing$, the retention contrast (§2), in the shared-controller study and the post-training study (§6, Appendix L)
cell	a (model, problem) pair; within one family, a problem
s_φ, r_φ	slope of φ's cell mean on screening fresh-draw success; $r_\varphi := s_\varphi/s_A$ (Appendix G)
LS, FS, LR; PRL, ORL	LoRA and full-parameter supervised fine-tuning, LoRA GRPO (§6); S2R process- and outcome-RL checkpoints (Appendix J)
$\mathcal{E}_A$, $\mathcal{E}_{D_\varnothing}$, $\mathcal{E}_{D_{\text{ret}}}$	end-to-end success: the model's own first attempt, then one verified retry episode under A, $D_\varnothing$ or D_{ret} (Table 24)

C SUPPORTING THEORY: FULL STATEMENTS AND PROOFS

The main text states Definition 1 and Lemma 1 in §2 and Proposition 1 in §4, where it uses the observables of Definition 2 below and cites Theorem 1, informally, and Proposition 2. This appendix, the idealized framework, proves them and states the supporting results in full.

Definition 2 (Retry policy and its observables)**.** A *retry policy* φ is a rule that continues from a failed state under an output-token ceiling b and ends on one graded answer Z_φ. Its *success* is $c_\varphi := \mathbb{P}(Z_\varphi = y^\star)$ and its *leave rate* is $g_\varphi := \mathbb{P}(Z_\varphi \neq y)$, read on the graded answer, so a policy that leaves the class and returns counts as staying. When $g_\varphi > 0$, its *changed-proposal excess* is $\xi_\varphi := c_\varphi/g_\varphi - v_y$.

Two scopes are used. In the *fixed-truth* scope $y^\star$ is a fixed element of $\mathcal{Y}$ and all statements are about the kernel $K(\cdot \,|\, y)$ on $\mathcal{Y}$ that one call of a policy induces from the failed state, with *changed-conditional law* $\check{K}(z \,|\, y) := K(z \,|\, y)/(1 - K(y \,|\, y))$ for $z \neq y$; in the *random-truth* scope the truth Θ is random, the sampler is calibrated ($\mathbb{P}(\Theta = z \,|\, X) = p_0(z)$), the rejected answer $Y \sim p_0$ is drawn independently of Θ given the context X, the verifier bit is $W := \mathbf{1}\{Y = \Theta\}$, and everything conditions on $\{W = 0, Y = y\}$.

C.1 THE LEDGER

Proof of Proposition 1. Fix a state on which A leaves the class ($g_A > 0$, as the proposition assumes) and write c_φ, g_φ for its observables. If $g_\varphi > 0$, the changed-proposal excess $\xi_\varphi = c_\varphi/g_\varphi - v_y$ of Definition 2 gives $c_\varphi = g_\varphi(v_y + \xi_\varphi)$ at once, and $P_\varphi = c_\varphi - g_\varphi v_y = g_\varphi \xi_\varphi$. For the identity, $v_y := c_A/g_A$ gives $c_A = g_A v_y$, hence

$$R_\varphi + P_\varphi = (g_\varphi - g_A)v_y + c_\varphi - g_\varphi v_y = c_\varphi - g_A v_y = c_\varphi - c_A,$$

with no constraint on g_φ, so the identity holds on every such state. If $g_\varphi = 0$ the graded answer lies in y's class with probability one; since $y^\star \neq y$ this forces $c_\varphi = 0$, so $P_\varphi = 0 - 0 \cdot v_y = 0$ and $R_\varphi = -g_A v_y = -c_A$, and $c_\varphi - c_A = -c_A$ as required. Taking $\varphi = A$ gives $P_A = c_A - g_A v_y = 0$ and $\xi_A = c_A/g_A - v_y = 0$. □

C.2 COPY INVARIANCE

A revision step admits a copy/work decomposition $K = \lambda \mathrm{Id} + (1 - \lambda)\tilde{K}$, the mover–stayer mixture of Blumen et al. (1955) and Goodman (1961); the copy (stayer) mass is not identifiable from the one-step kernel K, since for any $\lambda' \in [0, \min_y K(y \,|\, y)]$ the same kernel decomposes with copy mass λ'. Only quantities that do not depend on λ' can therefore be read from K.

Lemma 2 (Copy invariance)**.** *Let $K(y \,|\, y) < 1$ and $\check{K}(z \,|\, y) := K(z \,|\, y)/(1 - K(y \,|\, y))$ for $z \neq y$. For every admissible copy decomposition $\lambda \leq K(y \,|\, y)$, the changed-conditional law computed from $\tilde{K}$ coincides with the one computed from K.*

Proof. For $z \neq y$, $\tilde{K}(z \,|\, y) = K(z \,|\, y)/(1-\lambda)$ and $1 - \tilde{K}(y \,|\, y) = (1 - K(y \,|\, y))/(1-\lambda)$. Dividing, the factor $(1 - \lambda)$ cancels. □

C.3 THE LOCAL-ANCHORING BOUND

Call a step *locally anchored* at $y \neq y^\star$ if $K(\cdot \,|\, y) = \lambda \mathrm{Id} + (1 - \lambda)[(1 - \alpha_y)p_0 + \alpha_y \mu_y]$ for some anchoring strength $\alpha_y \in [0, 1]$ and anchoring law μ_y supported on $N[y] := \{y\} \cup N(y)$, for a neighbourhood $N(y)$ of y, with $y^\star \notin N[y]$: not knowing whether it is wrong, the step can copy, re-derive from the base law, or be pulled toward the previous answer, but the pull cannot point at $y^\star$. Write $m := \mu_y(y)$. Policies that reshape the whole law, such as a greedy or top-T decision rule, sit outside this family and are priced separately (Proposition 4).

Theorem 1 (Local-anchoring bound)**.** *Let the step be locally anchored at y, and let $q_0 > 0$, $p_0(y) < 1$ and $K(y \,|\, y) < 1$, so that the changed-conditional law $\check{K}(\cdot \,|\, y)$ is defined. Then $\check{K}(y^\star \,|\, y) \leq v_y$, with equality if and only if $\alpha_y = 0$ or $\mu_y(y) = 1$.*

Corollary 1. *If the step is locally anchored and its changed-proposal excess $\check{K}(y^\star \mid y) - v_y$ is negative at a state (for a policy whose graded answer is one draw of the step, this excess is ξ_φ), then $\alpha_y > 0$ and $\mu_y(y) < 1$: within this family, displaced mass lands near the previous answer without landing on it. Exact repetition of y moves the routing term, not the proposal term. The converse inference is not available: a step that moves the base mass of one wrong answer $w \neq y$ and half of $y^\star$'s onto an answer $w' \notin \{y, w, y^\star\}$, with $q_0 > 0$ and $p_0(w) > 0$, has a negative excess but no locally anchored form, since that form gives every $z \neq y$ at least $p_0(z)/q_0$ times the mass it gives $y^\star$.*

Proof of Theorem 1. Write $\alpha = \alpha_y$, $m = \mu_y(y)$. Then $K(y^\star \mid y) = (1-\lambda)(1-\alpha)q_0$ and $1 - K(y \mid y) = (1-\lambda)[1-(1-\alpha)p_0(y) - \alpha m]$, so

$$\check{K}(y^\star \mid y) = \frac{(1-\alpha)q_0}{1-(1-\alpha)p_0(y) - \alpha m}.$$

Both denominators are positive ($1 - K(y \mid y) > 0$ forces $\lambda < 1$ and a positive bracket, and $p_0(y) < 1$), so cross-multiplying against $q_0/(1-p_0(y))$ reduces the claimed inequality to $\alpha(1-m) \geq 0$, which always holds, with equality exactly when $\alpha = 0$ or $m = 1$. □

Proof of Corollary 1. A negative excess means $\check{K}(y^\star \mid y) < v_y$, i.e. strict inequality in Theorem 1, which by the equality condition forces $\alpha_y > 0$ and $m < 1$. Anchoring that lands exactly on y ($m = 1$) is absorbed into the copy atom by the conditioning of Lemma 2 and leaves $\xi = 0$; what it moves is g_φ, hence the routing term. □

C.4 The verdict-only restart

Proof of Lemma 1. Each raw draw is from p_0, and rejecting y's class until a draw is accepted returns one draw from $p_0(\cdot \mid \cdot \neq y) = \rho$. So $K(\cdot \mid y) = \rho$ has no diagonal mass, $K(y \mid y) = 0$, $\check{K} = K = \rho$ and $g_B = 1$. Hence $c_B = \rho(y^\star) = q_0/(1-p_0(y)) = v_y$ and $\xi_B = c_B/g_B - v_y = 0$. The excess over a fresh draw is $v_y - q_0 = q_0[1/(1-p_0(y)) - 1] = q_0 p_0(y)/(1-p_0(y))$, which is positive exactly when $q_0 > 0$ and $p_0(y) > 0$. In Eq. (2), $R_B = (1-g_A)v_y \geq 0$ and $P_B = c_B - 1 \cdot v_y = 0$. □

The finite-budget readings of B are in §5 and Appendix H, and the cap-hit readings in Appendix I.

C.5 The posterior theorems

Throughout this subsection the random-truth scope is in force and ρ is the reference law of Definition 1.

Theorem 2 (The reference is the posterior)**.** *Under calibration, for $z \neq y$, $\mathbb{P}(\Theta = z \mid X, Y{=}y, W{=}0) = p_0(z)/(1-p_0(y)) = \rho(z)$.*

Proof. Given X, conditional independence gives joint mass $\mathbb{P}(\Theta = z, Y = y) = p_0(z)p_0(y)$. Conditioning on $\{Y = y\} \cap \{\Theta \neq y\}$ gives numerator $p_0(z)p_0(y)\mathbf{1}\{z \neq y\}$ and denominator $p_0(y)(1-p_0(y))$; cancel $p_0(y)$. □

Theorem 3 (Value of the verdict at the realized truth)**.** *Conditional on the realized truth $\Theta = y^\star$, a sampler that draws from its coherent belief, given (X, Y) or given (X, Y, W), succeeds with probability q_0 without the verdict and $q_0/(1-p_0(y))$ with it, so the verdict is worth $q_0\, p_0(y)/(1-p_0(y))$, exactly the excess of c_B over q_0 in Lemma 1. Averaged over $\Theta \sim \rho$ the with-bit success is $\|\rho\|^2$ and the value of the verdict is $p_0(y)\,\|\rho\|^2$.*

Proof. Without the verdict the belief remains p_0, since the sampler's own draw carries no information about Θ, and success is q_0; with it the belief is ρ by Theorem 2 and success is $\rho(y^\star) = q_0/(1-p_0(y))$. Subtract. For the averaged version, $\mathbb{E}_{\Theta\sim\rho}[\rho(\Theta)] = \sum_z \rho(z)^2 = \|\rho\|^2$ and $\mathbb{E}_{\Theta\sim\rho}[p_0(\Theta)] = \sum_z \rho(z)p_0(z) = (1-p_0(y))\|\rho\|^2$. □

Theorem 4 (Beyond the verdict, within a declared class)**.** *Say a policy belongs to the class $\mathcal{B}$ of Bayes-rational, belief-consistent samplers if $\check{K}(\cdot \mid y) = \mathbb{P}(\Theta = \cdot \mid \mathcal{G})$ for some σ-algebra $\mathcal{G} \supseteq \sigma(X, Y, W)$, with the draw Z conditionally independent of Θ given $\mathcal{G}$. For any member, on $\{W = 0\}$,*

$$\mathbb{E}[\xi \mid X, Y{=}y, W{=}0] = \mathbb{E}\|\mathbb{P}(\Theta \mid \mathcal{G})\|^2 - \|\rho\|^2 \;\geq\; 0,$$

with equality if and only if $\mathcal{G}$ carries no information about Θ beyond $\sigma(X, Y, W)$; this is monotonicity of the expected quadratic score in the Blackwell order (Blackwell, 1953; DeGroot & Fienberg, 1983). The realized-truth version, which is the quantity estimated per state, also holds: conditional on $\Theta = y^\star$, $\mathbb{E}[\check{K}(y^\star \mid y) \mid \Theta{=}y^\star] - \rho(y^\star) = \mathrm{Var}[\mathbb{P}(\Theta{=}y^\star \mid \mathcal{G})]/\rho(y^\star) \geq 0$, with equality if and only if that posterior coordinate is almost surely constant. Both statements average over the signal carried by $\mathcal{G}$; on one fixed failed transcript a member of $\mathcal{B}$ can realize $\xi < 0$.

Proof. By the tower property, $\mathbb{E}[\check{K}(\Theta \mid y)] = \mathbb{E}[\sum_z \mathbf{1}\{\Theta = z\}\mathbb{P}(\Theta = z \mid \mathcal{G})] = \mathbb{E}\|\mathbb{P}(\Theta \mid \mathcal{G})\|^2$, while $\mathbb{E}[\rho(\Theta)] = \|\rho\|^2$. Coordinatewise conditional Jensen gives $\mathbb{E}[\mathbb{P}(\Theta = z \mid \mathcal{G})^2] \geq \rho(z)^2$; summing gives the inequality, with equality if and only if each posterior coordinate is almost surely constant, i.e. $\mathbb{P}(\Theta \mid \mathcal{G}) = \rho$ almost surely. For the realized-truth version write $U := \mathbb{P}(\Theta{=}y^\star \mid \mathcal{G})$; then $\mathbb{E}[U \mid \Theta{=}y^\star] = \mathbb{E}[U\mathbf{1}\{\Theta{=}y^\star\}]/\rho(y^\star) = \mathbb{E}[U^2]/\rho(y^\star)$ by the tower property, and $\mathbb{E}[U^2] = \rho(y^\star)^2 + \mathrm{Var}[U]$. □

Proposition 2 (A belief-consistent benchmark)**.** *If the truth Θ is random with a calibrated sampler, $\mathbb{P}(\Theta = z \mid X) = p_0(z)$, the rejected answer $Y \sim p_0$ is drawn independently of Θ given the context X, and the verdict is $W := \mathbf{1}\{Y = \Theta\}$, then $\mathbb{P}(\Theta = z \mid X, Y{=}y, W{=}0) = \rho(z)$ for $z \neq y$. Within the class of belief-consistent samplers, $\mathbb{E}[\xi \mid X, Y{=}y, W{=}0] \geq 0$, with equality exactly when the policy carries no information about the truth beyond the verdict.*

Proof of Proposition 2. The posterior statement is Theorem 2 and the sign statement is Theorem 4. Theorem 3 adds the success of sampling that posterior at the realized truth, which the single accepted draw of Lemma 1, the verdict-only reference, attains. □

D THE INFORMATION–POLICY DECOMPOSITION

All statements here live in the random-truth scope of Appendix C, with $I_0 := \sigma(X, Y, W)$, beliefs $\beta = \mathbb{P}(\Theta = \cdot \mid \mathcal{G})$ for some $\mathcal{G} \supseteq I_0$, and action law Π supported on $\{\neq y\}$. They bear on three points in the main text: Proposition 3 is the sense in which the exclusion is the minimal update (§7); Proposition 4, that a ξ level alone does not identify information, is why §5 reads the ledger as behavior and not as information use; and Proposition 5, proved here, is the false-verdict boundary of §7.

Theorem 5 (Information–policy decomposition)**.** *The population excess $\bar{\xi} := \mathbb{E}[\xi \mid I_0]$ satisfies $\bar{\xi} = \bar{\xi}_{\mathrm{info}} + \bar{\xi}_{\mathrm{pol}}$, where $\bar{\xi}_{\mathrm{info}} := \mathbb{E}\|\beta - \rho\|^2 \geq 0$ is the resolution term of the Brier decomposition (Murphy, 1973) and $\bar{\xi}_{\mathrm{pol}} := \mathbb{E}\langle \Pi - \beta, \beta \rangle$. With $\Pi = \beta$ (sampling one's own beliefs), $\bar{\xi}_{\mathrm{pol}} = 0$ and Theorem 4 is the special case; with $\beta = \rho$ (no extra information) the excess is pure policy, and sharpening and anchoring are its two signs.*

Proof. Since $Z \perp \Theta \mid \mathcal{G}$, $\mathbb{P}(Z = \Theta \mid \mathcal{G}) = \langle \Pi, \beta \rangle$, so $\bar{\xi} = \mathbb{E}\langle \Pi, \beta \rangle - \|\rho\|^2 = (\mathbb{E}\|\beta\|^2 - \|\rho\|^2) + \mathbb{E}\langle \Pi - \beta, \beta \rangle$. The tower property gives $\mathbb{E}[\beta \mid I_0] = \rho$, hence $\mathbb{E}\langle \beta, \rho \rangle = \|\rho\|^2$ and $\mathbb{E}\|\beta - \rho\|^2 = \mathbb{E}\|\beta\|^2 - \|\rho\|^2$. □

Proposition 3 (Exclusion is the I-projection)**.** *For $\mathcal{H} \subset \mathcal{Y}$ with $y \in \mathcal{H}$ and $p_0(\mathcal{H}) < 1$, the exclusion update $\rho_{\mathcal{H}} := p_0(\cdot \mid \cdot \notin \mathcal{H})$ satisfies $\rho_{\mathcal{H}} = \arg\min_{q:\, q(\mathcal{H})=0} \mathrm{KL}(q \,\|\, p_0)$, the I-projection of Csiszár (1975) onto $\{q(\mathcal{H}) = 0\}$, with minimum $-\log(1 - p_0(\mathcal{H}))$.*

Proof. For $q(\mathcal{H}) = 0$, $\mathrm{KL}(q\|p_0) = \mathrm{KL}(q\|\rho_{\mathcal{H}}) - \log(1 - p_0(\mathcal{H})) \geq -\log(1 - p_0(\mathcal{H}))$, with equality if and only if $q = \rho_{\mathcal{H}}$. □

Proposition 4 (Scalar non-identification). *For every $a \in [0, \max_z \rho(z) - \|\rho\|^2]$ there exist (i) a zero-information sharpener, $\beta = \rho$ and $\Pi = (1-t)\rho + t\,\delta_{\arg\max \rho}$ with $t = a/(\max_z \rho(z) - \|\rho\|^2)$, and (ii) a partial-revelation channel sampled faithfully, revealing Θ with probability $u = a/(1 - \|\rho\|^2)$ and nothing otherwise, with $\Pi = \beta$ (for $a = 0$ take $t = u = 0$); both have $\bar{\xi}$ exactly a. The interval has non-empty interior if and only if ρ is non-uniform on its support.*

Proof. For (i), $\beta = \rho$ gives $\bar{\xi}_{\text{info}} = 0$, so by Theorem 5 $\bar{\xi} = \bar{\xi}_{\text{pol}} = \langle \Pi - \rho, \rho \rangle = t(\max_z \rho(z) - \|\rho\|^2) = a$. For (ii), $\Pi = \beta$ gives $\bar{\xi}_{\text{pol}} = 0$ and $\bar{\xi} = \bar{\xi}_{\text{info}} = \mathbb{E}\|\beta - \rho\|^2 = u\,\mathbb{E}_{\Theta\sim\rho}\|\delta_\Theta - \rho\|^2 = u\,\mathbb{E}_{\Theta\sim\rho}[1 - 2\rho(\Theta) + \|\rho\|^2] = u(1 - \|\rho\|^2) = a$. Both require their parameter in $[0, 1]$, which holds because $\max_z \rho(z) - \|\rho\|^2 \le 1 - \|\rho\|^2$. Finally $\|\rho\|^2 = \sum_z \rho(z)^2 \le \max_z \rho(z)$ with equality exactly when $\rho(z) \in \{0, \max_w \rho(w)\}$ for every z, i.e. when ρ is uniform on its support. □

Proposition 5 (Noisy verifier: soft reweighting). *Let the verifier emit e about y with true-positive rate $\text{TPR} := \mathbb{P}(e{=}\text{wrong} \mid \Theta \neq y)$ and true-negative rate $\text{TNR} := \mathbb{P}(e{=}\text{correct} \mid \Theta = y)$, its error depending on Θ only through $\mathbf{1}\{\Theta = y\}$. On $e = \text{wrong}$ the coherent update is $\tilde{\rho}(z) \propto p_0(z)\,\text{TPR}$ for $z \neq y$ and $\tilde{\rho}(y) \propto p_0(y)(1 - \text{TNR})$: hard exclusion is the case $\text{TNR} = 1$, and a verifier that can reject a correct answer ($\text{TNR} < 1$) but is informative, $\text{TPR} + \text{TNR} > 1$,* down-weights *the refuted answer rather than masking it: the odds of y against any other answer are multiplied by $(1 - \text{TNR})/\text{TPR}$, which is below one exactly under that condition. A hard cumulative ban discards the posterior mass $\tilde{\rho}(y) = p_0(y)(1-\text{TNR})/[p_0(y)(1-\text{TNR}) + (1-p_0(y))\text{TPR}]$ that a correct but falsely rejected answer retains, and does so permanently, whereas a revision policy leaves such an answer only with its correct-start leave probability $\mathbb{P}(Z \neq y \mid y, \Theta{=}y)$, the correct-start analogue of g_φ.*

Proof of Proposition 5. Conditioning on $e = \text{wrong}$ and using that the verifier's error depends on Θ only through $\mathbf{1}\{\Theta = y\}$, $\mathbb{P}(e{=}\text{wrong} \mid \Theta = z) = \text{TPR}$ for $z \neq y$ and $= 1 - \text{TNR}$ for $z = y$. Bayes' rule with prior p_0 gives $\tilde{\rho}(z) \propto p_0(z)\text{TPR}$ for $z \neq y$ and $\tilde{\rho}(y) \propto p_0(y)(1 - \text{TNR})$, so the odds of y against any $z \neq y$ are the prior odds times $(1 - \text{TNR})/\text{TPR}$, and normalizing gives the stated $\tilde{\rho}(y)$. At $\text{TNR} = 1$ the second is zero and $\tilde{\rho} = \rho$, the law of the accepted draw in Lemma 1. Under a cumulative hard ban the mass $\tilde{\rho}(y)$ is removed from every later round, which is the stated loss. □

E Protocol details and provenance

States. For each screened problem the model's own clean-context draws are graded; a problem is eligible when its fresh-draw success lies in $[2/64, 62/64]$, and two of its wrong draws are frozen as starts. Confirmation problem sets were sealed before any policy ran and unsealed once, for the nine-policy collection. A state is (problem, start) and is run with eight episodes per (policy, ceiling). The starts are never re-drawn, so all policies of a (family, ceiling) run on the same states.

Episodes. A policy runs as a loop of calls against the tested model until the ceiling would be exceeded (fewer than 64 tokens remaining ends the loop), the 64-call cap is hit, or the policy's own rule stops it. Each call is capped at $\min(4{,}096, \text{remaining})$ output tokens, except at the 16k and 32k ceilings (the three confirmation regimes, the paired effort runs and the thinking switch), where the per-call cap equals the ceiling. All solver draws are billed, rejected and retried draws included; the prefix judge (a different family, ≤ 256 tokens) and the `sympy` fact checker are not.

Table 8: **Further studies** (§§2–6), in the notation of Tables 1 and 3; contrasts revision − restart and $D_{\text{ret}} - D_{\varnothing}$.

Study	Policies	Instruction	Told wrong	Controller
S2R checkpoints	revision; restart; fresh draw	revise; none; none	revision	restart: $\le 7 \times 4$k, no episode ceiling
Shared-controller study	$D_{\varnothing}$ (withheld), D_{ret}	revise ($D_{\varnothing}$: hidden)	–	32 calls, no forced acceptance, 8/16k
Post-training study	A; $D_{\varnothing}$, D_{ret}	none; as above	$D_{\varnothing}$, D_{ret}	seventh draw accepted

Controllers and prompts. Four controllers appear, and each bans one class only, the frozen start's. *Nine-policy collection:* a draw is accepted if it is complete and outside the banned class; a banned or truncated draw is redrawn and the seventh consecutive draw is accepted regardless (B_{budget} counts only banned draws toward the seven and never accepts a truncated one), and policies that stop at the ceiling keep drawing until it is spent. *Shared-controller study:* one output-token budget of 8k (16k) per episode, each call capped at what remains, at most 32 calls, no forced acceptance; a truncated call, or a complete one without a boxed answer, ends the episode as a failure, a complete draw in the banned class is rejected and billed, and any other complete draw is accepted. *S2R study* (Appendix J): an earlier controller that accepts any draw whose parsed class is not banned, truncated or unparseable included, keeps the seventh draw after six rejections, and caps each draw at $4{,}096$ tokens ($16{,}384$ in the budget check) with no episode ceiling; its fresh draw is a single draw. *Post-training study:* the nine-policy rule for B at an 8k ceiling, each call capped at $\min(4{,}096, \text{remaining})$, with a final answer that carries no box scored as wrong. The S2R study's verdict-only restart, like the nine-policy B, is prompted with the bare problem. In the shared-controller study and the post-training study the prompts are those of §2, with the line "[Previous solution withheld.]" as $D_\varnothing$'s assistant turn. The post-training A is the bare problem, so there $D_\varnothing$ is drawn inside a revision prompt and A is not. The prompt has a price of its own on models not trained on it: in an offline analysis specified after the sealed test was read, $D_\varnothing - A$ is $-.036$ $[-.059, -.014]$ on the base model and $-.048$ $[-.073, -.024]$ for the GRPO model (LR), against $+.021$ $[-.003, +.045]$ and $+.017$ $[-.002, +.035]$ for the two models fine-tuned on D_{ret} prompts (levels in Table 5; problem-paired, 95%). Pricing the kept transcript against $D_\varnothing$ and not against A keeps that price out of the retention contrast.

Policies. The revision instruction reads "Here is your previous solution. Review it carefully, find any error, and produce an improved final solution."; its hidden variant, used where the previous solution is not shown, reads "You produced a previous solution to this problem, but it is not shown to you here. Review the problem carefully, find any error in your approach, and produce an improved final solution." Only B, B_{budget} and E exclude the start's class, and only the first calls of F and H state that the previous answer is incorrect. A: one fresh clean-context draw, retried until a call completes. B, the verdict-only restart: A's prompt under the nine-policy controller, banning the start's answer class. G: A's mechanism with no exclusion and no early stop, drawing fresh samples until the output-token ceiling is spent, the last complete draw graded. B_{budget}: G's loop with B's exclusion applied to every draw, graded on its last accepted draw; Llama-3.1-8B and Gemma-3-12B only. E: a revision chain each of whose steps shows a length-matched, content-free stand-in in place of the whole previous attempt (reasoning and answer) under the hidden revision instruction, plus B's exclusion loop; since its controller excludes y's class, its admissible set is B's and its ledger reference is ρ. C_{last}: a revision chain from the previous output only, under the revision instruction. C_{full}: the registered continuation policy, an author-designed full-history continuation baseline inspired by the continuation stage of Bilal et al. (2026), whose continuation conditions on the last corrected trace, allows rewriting and is followed by critique and correction; ours shows every earlier output and forbids restarting, and its wording is ours, since that paper publishes no prompt text. Each call shows the problem, then every previous output, the frozen start first, as "[Attempt 1]", "[Attempt 2]", …, in one assistant turn, then the instruction "Continue the solution above from where it stops. Do not start over and do not repeat the work already shown; carry it forward to a complete final answer."; it states no verdict, excludes no class and makes no critique or correction call. D: each step re-judges the previous step's output with a cross-family locator and continues from the located prefix, with no instruction; a judge failure degrades that step to C_{last} (judge-failure rate 11–13% with the Llama solver, 31% with the Gemma solver). F: intermediate equalities are extracted from the failed attempt once and the first six are checked with `sympy` for validity and for dependence on established symbols; at most three that pass, within 48 tokens, are shown on the first call in a review note that also states that the previous answer is incorrect, over E's content-free stand-in, the failed text not shown; later steps are plain C_{last}. H: the same memo appended, on the first call, to the failed transcript shown verbatim under the revision instruction; later steps are plain C_{last}.

Extractor and checker. A fact is an explicit intermediate equality in the failed attempt's text, in the form $\text{lhs} = \text{rhs}$, with its source line; chained equalities are split into adjacent pairs. Lines carrying the final answer (`\boxed`, "the answer is"), trivial identities and lines over 120 characters are dropped. A fact is *valid* when both sides parse and `sympy` simplifies their difference to zero or matches them numerically to 10^{-9}. A fact is *admissible* when in addition its free symbols lie in

the union of the problem's symbols and the symbols defined by earlier valid facts. A gold guard removes any fact whose right-hand side equals the gold answer; the extractor, the checker and the memo prompt were frozen before any memo was built, and the checker was re-frozen once, before the Qwen3-32B, Gemma-4-31B and memo positive-control memos (Appendix G).

Confirmation regimes. Gemma-4-31B-it ran with thinking on at a 16k ceiling, the lowest at which the fresh draw's truncation on the screening draws is at most 5% (97% and 86% of draws would truncate at a 4k or 8k per-call cap, 4.8% at 16k); the realized fresh-draw truncation is 6.1%. Its set closed at 71 problems after three screening rounds. The thought channel is split from the answer at the serving boundary, stored, billed as output tokens, and never carried into a later call; the history policies carry only the visible answer text. gpt-oss-120b ran at reasoning effort low and high, the same weights under two settings; low still reasons briefly. Each setting has its own problem set drawn from a 213-problem competition set: at low effort more were eligible and the set was capped at 80, at high effort 61 were eligible; contrasts are paired within a setting. Truncation: A ends at the cap in 0.1% of episodes at low effort and 23.1% at high, B in 0.0% and 25.5%, C_{full} in 0.0% and 6.5%; C_{full} makes 59.6 calls per episode at low effort and 16.1 at high. The paired effort runs (Table 14) put each effort on the other setting's problems, on seed blocks disjoint from every earlier run. Qwen3-32B's thinking switch ran three settings on one problem set — thinking on at 16k, off at 16k, and the existing off setting at 4k.

Inference. A cell's two starts and eight episodes average within it; on the collection analyses the t interval agrees with the cell bootstrap to within .008, and family-wise control is in Appendix I.

Seeding. Every (policy, ceiling, family) block draws its generation seeds from a registered, disjoint block; a pre-launch assertion scans the stored run records to confirm that no record already carries a seed of the new block, and a check after each condition confirms that none of its records carries a seed outside it. The earliest nine-policy offline analyses were bootstrapped with one shared random stream across groups, which makes a group's interval depend on which other groups were resampled in the same batch; they were re-run with each (family, policy, ceiling, contrast) group seeded independently, and the max-T resample, one joint stream by design, under a new seed. Point estimates are unchanged to the printed digit, the largest absolute interval movement is .0067, the max-T threshold moves from 3.2402 to 3.18, and no registered decision changes. The collection analyses draw all the contrasts of one analysis from one stream.

Answer equivalence and grading. Mathematics answers are normalized symbolically (`sympy`, floats to rationals) and compared as equivalence classes, not strings; the same normalizer defines the controller's banned class. A stratified human audit of the answer-equivalence check agreed on $54/60$ items; all six disagreements fall in the registered string-fallback subset, and there are zero reverse-direction errors. That fallback, reached whenever symbolic parsing fails, judged surface-form variants of a correct answer wrong and covers $59/500$ (11.8%) of the MATH-500 gold values; it was replaced by a canonical normalizer and the earlier archive was re-judged. The collections reported in this paper ran after that replacement, with the canonical normalizer set by each runner at launch and asserted on the node, by a version check in the collection runners and by a start-up check in the post-training and memo positive-control scripts: screening for the frozen wrong starts and the controller's accept and reject decisions used it. Appendix K reports a later scan with wider surface normalizations and the results with and without the starts it flags.

Episode table and cross-checks. The offline analyses run on a flattened episode table built from the raw run records: 69,120 rows for the Llama and Gemma nine-policy collection (54 (policy, ceiling, family) blocks $\times$ 1,280 episodes; B_{budget}, collected later, is not in it) and 12,800 rows for Qwen. Every block's cell mean is checked against the collection results digit by digit before any analysis runs, and the ledger's additivity gap (Proposition 1) is computed on every cell, with a maximum absolute value of 0.

Serving. All inference is served locally with vLLM (Kwon et al., 2023), one model per allocation and four servers per run, except the nine-policy prefix judges (two models, one server each, on one L40S allocation), and the served model identity is checked against the registered identifier before a run starts. *GH200:* the nine-policy collection's solver draws, the three confirmation regimes, the paired effort runs and the thinking switch; Gemma-3-12B's shared-controller policies; the transcript-part and other-problem conditions for Gemma-3-12B and Qwen3-8B; the S2R common-start run; the post-training sealed test and self-error check; the first calls the memo positive control is read

against; the selection and test of the two further LS seeds and the mixed-data control; the end-to-end study. *H200:* Qwen3-32B's shared-controller policies; the transcript-part and other-problem conditions for Llama-3.1-8B and Qwen3-32B (on Llama-3.1-8B they re-collected both policies on one node the same day); the 16k revision collection of Appendix J, its 16k verdict-only restart and fresh-draw policies and its control against the earlier run; the post-training checkpoint-selection pass; the code collection; the memo positive control; the training of the two further LS seeds and the mixed-data control. *H100:* Qwen3-8B's shared-controller policies, so its comparisons with its transcript-part and other-problem conditions cross types; the serving-precision control, on H100 only. The Llama-3.1-8B shared-controller study collected its policies in problem chunks on GH200 and H200 nodes without balancing the two policies across node types (Appendix K).

Provenance. Every reported number is read from a stored analysis output or computed by arithmetic on such outputs, except three values from overwritten earlier runs of the nine-policy offline analyses (above and Appendix I) and the counts from two human readings (the audit of the answer-equivalence check and the reading of kept facts), which are taken from the dated lab records. Protocol figures and offline checks first taken from those records (screening truncation shares, judge-failure rates, the adapter's training-time KL estimate, the content factorial's prompt-token totals, the code study's lenient first-call rate, and the gpt-oss-120b row with its flagged problem removed) were recomputed from the stored run records before submission. The nine-policy numbers and tables are emitted by generator scripts over the registered analyses, and the figure scripts refuse to run if any input checksum differs from the recorded one; the shared-controller analysis script was hash-recorded at each revision, and the post-training analysis script was frozen before its training data existed. The answer-normalization scan over all start sets and the check that drops the affected Qwen3 problems were run by ad hoc scripts kept with their outputs in the run records.

F What each contrast estimates

Every retry contrast in this paper except the end-to-end success of Table 24 and the classical statistic of Appendix J conditions on a failure state, and the contrasts differ in which failure states they condition on, at what point in the episode they are read, and whether they were registered before their data. Table 9 records these differences contrast by contrast. In its status column, "margin offline" and "2×2 offline" name the part of a row that is an offline analysis, and post hoc marks an analysis added after its data had been read; offline analyses specified after the results they re-examine are marked where they appear.

Table 9: What each contrast estimates. "Same states": whether both sides condition on the same frozen wrong starts; "measured at": the point of the episode where the statistic is taken; "status": reg. (registered before its data), offline (an analysis of collected episodes) or expl. (exploratory); deviations in Appendix G.

Tag (where)	Start sample	Same states	Measured at	Identifies	Status
CTRL (Figure 1)	one problem set per model, two wrong starts per problem; on Llama the two no-hint conditions of four	yes	graded episode end under the 32-call controller	$D_{\text{ret}} - D_{\varnothing}$, with the one-line withholding notice in the transcript's place	reg.; Llama discovery, others replication; 97 post hoc
NINE (Table 10, Figure 2)	one frozen set per family, two wrong starts per problem, shared by the family's policies	yes	graded episode end; no completed call is a failure	policy value at a matched output-token ceiling with all solver draws billed	reg.
FIRST (Table 13)	the NINE episodes	yes	first call the policy's own rule accepts, and the difference to the end	how much of the ladder exists before any continuation, and what the chain adds	offline
VSB (Table 12)	the NINE states of the seven (family, ceiling) settings where B ran	yes	graded episode end; per state for the split	$C_{\text{full}} - B$, and its split against B's measured routing and proposal terms	offline
THINK (Table 4, Appendix G)	one Qwen3-32B problem set, three settings	yes	graded episode end	the paired difference of differences across the thinking switch at a fixed ceiling	reg.
CONF (Table 4)	one problem set per confirmation regime, sealed before the policies ran	yes within a regime	graded episode end	the two endpoints under thinking and under two reasoning efforts	reg.
EFF (Table 14)	each effort's own problem set, run at both efforts on disjoint seed blocks	yes within a set	graded episode end	the paired difference of differences of the policy-minus-A contrasts ($B - A$, $C_{\text{full}} - A$) across the effort switch	reg.
FACT (Table 15)	a separate development set restricted to states carrying a checkable equality	yes	single draw (shown conditions) or accepted draw (hidden conditions)	content (source-attempt versus other-problem checked facts) and display; the first-call 2×2 on the NINE episodes as held-out replication	reg.; 2×2 offline
MEMO (Appendix H)	the NINE starts of problems with a checked reference-solution fact (66 Llama, 63 Gemma)	yes	first call at 4k	a problem's own against another problem's reference facts in the memo slot, and each against an empty memo	reg.
STOP (Appendix H)	the NINE set	yes	graded episode end	the stopping rule alone ($G - A$) and the exclusion alone ($B_{\text{budget}} - G$)	reg.
LEDGER (Figure 3, Table 11)	the NINE episodes	yes	per state, from the fresh draw's eight episodes; a 63-draw screening pool as a second reference (offline)	the split of Δc into routing and proposal (Proposition 1)	expl.
SHOWN (Tables 18, 19)	the CTRL problems with two boxed starts (130, 85, 100; 99 after one removal)	yes within a model	graded episode end and first call under the CTRL controller	the reasoning alone, the answer added back and the answer alone, at one instruction and a matched prompt length	reg.; Llama no-residual return post hoc
OTHER (Table 20)	the SHOWN states, the retained turn holding another problem's redacted failed transcript	yes	graded episode end and first call	the own against another problem's transcript, and the latter against a same-day placeholder	reg.
ENF (Tables 2, 21)	the CTRL episodes	yes	first call, before the controller acts; the calls after it; terminal state	$D_{\text{ret}} - D_{\varnothing}$ without enforcing the rejection, what enforcing it recovers, and where the loss lands	offline
S2R (Figure 4, Table 16)	2×80 frozen wrong starts of the untrained base, common to three checkpoints; each checkpoint's own 64-draw pool for the margin	yes (common starts); no (margin)	one revision call, or the restart's accepted draw	revision against the same checkpoint's verdict-only restart and fresh draw; each checkpoint's margin against its own verdict-only reference	reg.; margin offline
POST (Tables 5, 22)	a sealed test set of 256 problems, common to all models	yes	graded episode end under the nine-policy controller (seventh draw accepted)	$J = D_{\text{ret}} - D_{\varnothing}$ per model and each policy's change against the base; LS seeds, mixed data, serving precision	reg.
SELF (Table 23)	each model's own wrong starts, generated by that model	no, by construction	graded episode end	J on each model's own error population	reg.
E2E (Table 24)	400 test-split problems no failure filter touched; each model's own first attempts	yes within a model	first attempt, then one verified retry episode	$\mathcal{E}_{D_{\text{ret}}} - \mathcal{E}_{D_{\varnothing}}$ and $\mathcal{E}_{D_{\varnothing}} - \mathcal{E}_{A}$ within a model (LS: first seed); each policy's change against the base	reg.
CODE (Appendix K)	150 LiveCodeBench problems, two wrong fenced programs each	yes	graded episode end, sandboxed tests	$D_{\text{ret}} - D_{\varnothing}$ on code	reg.; lenient post hoc

G PREREGISTRATION AND DEVIATIONS

Every collection ran under a frozen, timestamped protocol specifying conditions, seeds, problem lists, estimators or analysis scripts, and per-prediction decision rules, with both directions declared publishable in advance, except that the nine-policy collection registered its effect sizes and criteria as planning targets; post-freeze changes exist only as append-only revision records, each timestamped and approved, some after records under them existed (a field rename in the nine-policy collector, the thinking switch's added criteria). The timestamps are the records' own: of the collections up to the post-training study, only the thinking switch and the second paired effort run were committed to version control before their data. Offline analyses are descriptive; the text marks those specified after the results they re-examine, and the parts added after an offline analysis's first run are listed below.

What was registered. The nine-policy protocol and its confirmation sets (frozen before any policy ran); policy value at a matched ceiling as the primary quantity, with the sample size powered on $B - A$ and $F - A$, and the cross-family replication rule; the exclusion $\times$ stopping 2×2; the content $\times$ display factorial with its decision rule (source-attempt facts $=$ other-problem facts $\Rightarrow$ change the narrative, no router); the third family at the 4k ceiling with three criteria; and the three further inference regimes with three criteria fixed before their results ($C_{\text{full}} - A$ below zero with the interval excluding it; $B - A$ not below zero; the sign of $F - A$ where F ran). The shared-controller study was collected as a four-condition transcript $\times$ hint design whose registered target quantity was the interaction; the no-hint retention contrast was one of the five contrasts its frozen readout reports, each with a Bonferroni-corrected interval, not the target, and had been read on a 16- and then a 32-problem probe before the powered collection was sized (Appendix K). The post-training study wrote its result rows into a frozen analysis script and recorded its hash before the training data existed, and its sealed test set was not run until the checkpoint selection file was written and its hash recorded. The later collections — the three further shared-controller models, the transcript-part conditions in three registrations, the other-problem transcript, the code domain, the memo positive control, the collected 16k S2R policies, two further LS seeds, the mixed-data and serving-precision controls and the end-to-end study — each registered conditions, contrasts and decision rules before their own data, the last four after the sealed test had been read.

The thinking switch, in full. Six criteria were fixed before any of its results. Writing s_φ for the slope of policy φ's cell mean on the set's screening fresh-draw success (a 64-draw index computed from the set's registered source, independent of policy A) and $r_\varphi := s_\varphi / s_A$: (i) a paired rise in $r_{C_{\text{full}}}$ with the interval excluding zero; (ii) r_B near 1 in both settings; (iii) C_{full} below A with $B - A$ not below zero with thinking on; (iv) no detected ceiling effect between the two thinking-off settings; and, added by a timestamped revision while the first of its six conditions was being collected, before any results, (v) a paired rise in C_{full}'s own slope with the interval excluding zero and (vi) the upper end of the difference of differences at or above zero. The mechanism claim would enter the paper only on (i), (iii), (v) and (vi) together. Measured: (i) $+2.190$ $[+.955, +11.016]$, (ii) r_B .91 and .95, (iv) $+.020$ $[-.049, +.082]$, (vi) $+.332$ $[+.228, +.430]$ all held; (v) did not ($+.127$ $[-.126, +.332]$); and the $B - A$ half of (iii) missed by a point estimate of $-.005$ $[-.020, +.012]$. The reading in §5 is therefore descriptive.

Deviations in the original collections. Qwen3-32B's 2k and 8k rows fall outside the original single-ceiling registration: C_{full} and F ran there under a conditional pre-approval issued before the 4k results were read, and A at those ceilings, which their contrasts need, was added after those results; B has no such extension. The Gemma-4-31B set closed at 71 problems against a registered 80 and was accepted without a fourth screening round, and the gpt-oss-120b high-effort set, under the same criteria, closed at 61 and was accepted on that precedent. A second thinking model served through a vendor API was registered as a confirmation setting and abandoned during screening on throughput; no number from it enters this paper. Policy B_{budget} was also registered for Qwen3-32B, to be read only against A and, at 4k, B, since Qwen3-32B does not run G; it was not run, because without G it completes no cell of the exclusion $\times$ stopping 2×2. The content $\times$ display factorial reached its registered "no router" branch; a router was nevertheless trained and tested afterwards on a separate approval, did not beat the best fixed policy, and enters no number here. The fact checker was re-frozen once, with a guard against oversized expressions, after the Llama-3.1-8B and Gemma-3-12B memos were built; the Qwen3-32B and Gemma-4-31B collections and the memo positive control ran on the revised checker and on a later build of the grader. gpt-oss-120b at high

effort exceeds the registered 5% truncation limit for the fresh draw (23.1%); the exception was decided before those policies ran, and a finished-episode reading specified after their results gives $B - A$ $+.063$ $[+.023, +.105]$ over 59 cells and $C_{\text{full}} - A$ $-.208$ $[-.282, -.133]$ over 60. The paired effort run registered its two differences of differences and the rule that both excluding zero in the narrowing direction would put the narrowing in the abstract; both held on the first set, the mirror run was registered with the same contrasts and rule, its whole-history difference of differences reversed sign, and the abstract sentence was withdrawn. The ledger of §4 fixed its estimators and criterion before any ledger number was computed but after the aggregate contrasts had been read, and its population-wide routing and proposal terms replaced a first version over surviving states after that version's numbers had been read; its criterion (content and history policies with ξ_φ not separated from zero and $|\xi_\varphi| \leq .05$, with at least two thirds of the gap carried by routing) fails on Llama-3.1-8B for every content and history policy at every ceiling, for C_{full} on Qwen at all three ceilings and on Gemma-3-12B at 4k (routing share passes, ξ_φ does not), and for E on Gemma-3-12B at 4k (interval excludes zero); row values are in Table 11. The Westfall–Young adjustment was specified after the primary results. The shared-controller study's registered hint check (the hint raising $D_\varnothing$ on 64 development problems, 95% lower bound above zero) passed on Llama-3.1-8B; its registered interaction with the transcript was collected and read and is not detected ($-.007$ $[-.046, +.034]$ over the 161 problems; all four conditions in Table 17), and the hint conditions were dropped from the main text after those results; an earlier two-hint construction failed its stop criterion (bar set after the first round) and was not used. The study's 161-problem result is a merged analysis of the 64 used to evaluate the hint check and 97 added afterwards, and the added 97 are reported separately, a split made after the merged results (Appendix K). In the post-training study the GRPO model had a registered "has it left the base" check, at least half of 128 greedy outputs differing from the base's; it read $.578$, so the model entered the test as registered, and a base-against-base control added after the check was specified reads $.523$, so the check does not resolve whether the adapter moved (Appendix L). The checkpoint-selection pass, registered for the GH200 nodes of the test, was moved to H200 nodes by a timestamped revision before it ran.

Deviations in the later collections. The three further shared-controller models were registered at 8k with a truncation pilot whose failure stopped the collection, Gemma-3-12B and Qwen3-32B on existing 80-problem sets. Before any of their episodes those two sets were replaced by each model's own screen on the 700-problem candidate set, which the frozen collector reads, and before either Qwen3 pilot ran the response to a failed pilot was changed to moving both policies to 16k, with their inclusion ruled on after collection; both Qwen3 pilots failed (Qwen3-8B $.39$, Qwen3-32B $.31$) and both rows are reported. Gemma-3-12B passed ($.023$); its 16k pair was added after its 8k results and is a sensitivity reading. The first transcript-part registration was written for all 161 problems and ran on the 130 whose starts carry a boxed answer, a restriction fixed after a smoke run and before any main episode; its reproduction check was re-specified then on the same 130, and the value printed is on those 130; its smoke stop rule on unparseable finals failed ($.031$, $.109$, $.208$) and was replaced before the main episodes by a rule on unparseable finals the grader scores correct. Its primary contrast is not detected, and the reading registered for that outcome (that the loss comes with the answer) is not used, because the answer-only condition is not a clean removal of the reasoning (read after collection). The redaction registration relaxed a smoke-test criterion that failed on one unboxed correct answer per redacted condition and added a lenient analysis; its residual-exposure split was registered for success, the first-call return within it was added after the first results, and the placeholder against $D_\varnothing$ is not a registered contrast. The three-model registration removed one Qwen3-32B problem from all four conditions after its smoke test, before any main episode; its prompt lengths are matched, not equal. The other-problem registration fixed a reading for a negative and for a null primary contrast: Llama-3.1-8B met the first (the loss is specific to the own transcript), Gemma-3-12B met the second (any failed transcript carries it) and the two Qwen3 models fell outside both; on Gemma-3-12B and the two Qwen3 models the registered reading is not applied, on a diagnostic written after the first two results; the donor-length sensitivity was added after the first results. By its lab record, which version control does not corroborate, the split of Table 2 fixed its tables and predictions before any split was computed; after its first run the parse flag was switched to the controller's own, the budget column was formed by merging a residual state, and the per-start correct-when-leaving contrast was added. The code collection lost its registered H100 share to a device fault before any record and ran all 150 problems on H200; its parse rule was found after the results not to be policy-neutral, its truncation pilot had run D_{ret} only, and its lenient analysis is post hoc. The memo positive control ran Gemma on the H200 node registered for

Llama, decided before any solver call; on Llama its primary contrast excluded zero, but the reading registered for that outcome presumed a gain of the problem's own reference facts over the empty memo, which is not detected, so the memo-slot reading of Appendix H was settled after the results; the first-call $F - E$ replication (Appendix H) was registered with this control. The collected 16k S2R policies started before the archive-wide stage of their seed-collision check finished; that stage flagged the block, and a search after collection found the flagged seeds only in the collection's own canary episodes. For the third LS seed the rule as written (a later checkpoint replaces an earlier one when higher by at least one validation episode) selects the 100% checkpoint, one episode ahead (219 against 218 of 512); the frozen selection code compared accuracies rounded to four decimals and kept the 50% checkpoint, which is the one tested, and the 100% checkpoint was not tested. The mixed-data control's gap between its best tested policy and the base's fresh draw was read outside its registration. The serving-precision control, registered for GH200, moved to one H100 node before any episode, and a bf16 configuration carrying the LoRA serving flags was added then. The end-to-end study, registered for H200, moved to the GH200 nodes of the sealed test before any record. The analysis scripts of the collected 16k S2R policies, the further LS seeds and the serving-precision control were written after collection, to estimators, intervals and decision rules registered before it. The ledger against B, the independent reference for v_y, the max-T index check, the answer-normalization scan and the post-training $D_\varnothing - A$ are offline analyses specified after the results they re-examine.

H THE NINE-POLICY COLLECTION: FULL RESULTS

Table 10 gives the nine-policy collection's contrasts against A. On Gemma-4-31B (16k, 71 cells) the fresh draw is .676, $G - A$ is $-.007$ $[-.033, +.018]$ and $F - A$ is $-.035$ $[-.070, -.003]$; gpt-oss-120b's fresh draw is .505 at low and .612 at high effort; the endpoints are in Table 4. Table 11 gives the ledger for every policy but B_{budget}, which was collected after the ledger had been computed, with v_y estimated from the fresh draw's eight episodes per state (q_0 = its success, $p_0(y)$ = its rate of drawing the start's class). States with $g_A = 0$ carry no v_y and are dropped for every policy. Per state the routing term is $(g_\varphi - g_A)v_y$ and the proposal term $c_\varphi - g_\varphi v_y$; the fresh draw's own proposal term is zero, so Δc equals their sum on every state (Proposition 1), and states where the policy never leaves the class stay in with a routing term of $-g_A v_y$. Cells average their states and the routing share is the routing mean over Δc where $|\Delta c| \geq .02$; Table 12 pairs that ledger with B on the same states. B's own proposal term at 4k is $-.022$ $[-.052, +.007]$, $-.001$ $[-.032, +.030]$ and $-.005$ $[-.028, +.018]$ on Llama, Gemma and Qwen. Table 13 gives the first-call reading at every ceiling, Table 14 the two paired effort runs, and Table 15 the checked-fragment factorial, described with the rest of that material below. Intervals for values §5 and §4 give as points: at 4k, $C_{\text{full}} - A$ is $-.334$ $[-.398, -.270]$ on Llama, $-.390$ $[-.477, -.311]$ on Gemma and $-.472$ $[-.546, -.394]$ on Qwen, and at its first call $-.395$ $[-.466, -.328]$ on Llama; on Llama at 4k, $E - A$ is $-.055$ $[-.093, -.018]$, $F - A$ $-.136$ $[-.174, -.098]$, $D - A$ $-.151$ $[-.189, -.110]$ and $C_{\text{last}} - A$ $-.198$ $[-.245, -.152]$; on Gemma at 4k, $F - A$ is $+.005$ $[-.032, +.044]$ and $D - A$ $-.038$ $[-.071, -.004]$; on Qwen, $F - A$ is $+.006$ $[-.027, +.038]$, $+.019$ $[-.013, +.050]$ and $+.016$ $[-.011, +.044]$ at 2k, 4k and 8k; on Qwen3-32B's thinking-switch problems at 16k, $C_{\text{full}} - A$ with thinking off is $-.458$ $[-.535, -.383]$; and at 4k C_{full}'s routing term is $-.233$ $[-.289, -.180]$, $-.371$ $[-.442, -.303]$ and $-.445$ $[-.512, -.377]$ and its proposal term $-.101$ $[-.140, -.061]$, $-.020$ $[-.051, +.012]$ and $-.029$ $[-.060, -.002]$ on Llama, Gemma and Qwen. The verdict-only restart leaves the rejected class at .991/.995/.996 on Llama, .920/.959/.988 on Gemma and .980 on Qwen; the whole history, told to continue the attempts shown, leaves far less often (Table 11). At 4k the whole history, which runs to the ceiling by rule, bills 4,094 and 4,090 output tokens per episode on Llama and Gemma against the fresh draw's 771 and 1,324 (episode table, Appendix E). The restart's readings in §5 are of a real controller, which redraws a banned or truncated draw only up to a cap and within the ceiling, not of the idealized single draw of Lemma 1 whose success is v_y.

Exclusion and stopping. Crossing the ban with the stopping rule (the fresh draw and the restart, each stopped at its first accepted draw or run to the ceiling) gives a stopping effect, fresh draws to the ceiling minus the fresh draw ($G - A$), whose interval contains zero in all 6 settings, and an exclusion effect, the restart to the ceiling minus fresh draws to the ceiling ($B_{\text{budget}} - G$), of the same sign in all 6, from $+.02$ to $+.08$, excluding zero in 3.

Table 10: **Nine policies, three families, three ceilings.** Row A: fresh-draw success (cell mean). Other rows: policy minus A at the same ceiling (80 cells $\times$ 2 starts $\times$ 8 episodes; seeded cell bootstrap). Bold: 95% interval excludes zero; ‡: survives the Westfall–Young adjustment over the 48 Llama/Gemma contrasts (B_{budget} is outside that adjustment set). Qwen3-32B ran A, B (4k only), F and C_{full}. Figure 2 is the 4k column with intervals.

		Llama-3.1-8B			Gemma-3-12B			Qwen3-32B		
Policy	keeps	2k	4k	8k	2k	4k	8k	2k	4k	8k
A	nothing (one fresh draw)	.561	.572	.578	.577	.630	.616	.638	.687	.705
B	exclusion of y's class only	**+.035**	+.007	+.024	**+.039**	**+.055**	**+.073**‡	–	**+.039**	–
G	nothing; budget spent on fresh work	−.021	−.011	−.004	+.002	+.006	+.001	–	–	–
B_{budget}	exclusion; budget spent on fresh work	−.003	+.009	+.034	+.020	**+.036**	**+.079**	–	–	–
E	content-free stand-in; exclusion	**−.044**	**−.055**	**−.043**	−.022	**−.044**	−.020	–	–	–
F	stand-in + checked facts, then C_{last}	**−.125**‡	**−.136**‡	**−.157**‡	−.007	+.005	+.033	+.006	+.019	+.016
D	judge-located correct prefix	**−.143**‡	**−.151**‡	**−.138**‡	**−.091**‡	**−.038**	+.008	–	–	–
C_{last}	last output, answer shown	**−.216**‡	**−.198**‡	**−.193**‡	**−.248**‡	**−.150**‡	−.052	–	–	–
H	whole transcript + checked facts	**−.221**‡	**−.239**‡	**−.176**‡	**−.198**‡	**−.119**‡	−.034	–	–	–
C_{full}	whole answer history	**−.334**‡	**−.334**‡	**−.341**‡	**−.395**‡	**−.390**‡	**−.338**‡	**−.459**	**−.472**	**−.479**

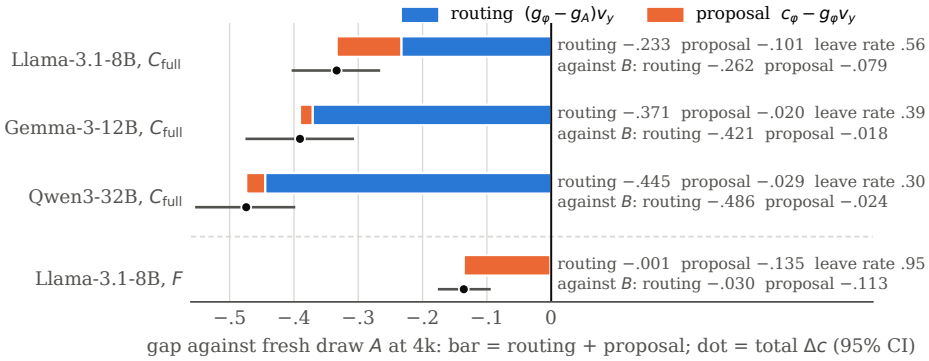

Figure 3: **The gap against a fresh draw, split into routing and proposal** (4k ceiling; Proposition 1). Bars: the two terms; dot: the gap Δc with its problem-level bootstrap 95% interval; right: the two terms and the leave rate g_φ, and below them the same two terms against the verdict-only restart B (intervals for the whole history in Table 12). C_{full}: whole history; F: checked facts; A: fresh draw. Every policy but B_{budget} at every ceiling, and the fresh draw's leave rates, in Table 11.

An independent reference for v_y. The ledger's v_y, taken from the fresh draw's episodes, shares noise with c_A and g_A and is undefined on 0 to 2 of the 160 states in the nine (family, ceiling) settings. The screening pool from which the starts were drawn gives a reference the policies do not share: 64 fresh draws per problem from the same model and prompt, cut to each ceiling by output length, with the draw that became the state's start removed (63 per state; all 80 problems for Llama-3.1-8B and Qwen, 34 of 80 for Gemma-3-12B, whose remaining pools were not retained). With an independent reference the fresh draw's own proposal term is no longer zero by construction, so the identity reads $c_\varphi - c_A = (g_\varphi - g_A)v_y + (P_\varphi - P_A)$, and P_A is a calibration check. Criteria for the comparison were fixed before the pool was read. P_A has an interval containing zero in all nine settings (−.016 to +.010). On the states both references cover, C_{full}'s routing share moves by at most .041 in any family and ceiling, and all 21 (family, ceiling, policy) proposal terms whose interval excludes zero under the fresh draw's reference keep their sign and still exclude zero, among them C_{full} on Llama at −.099 [−.138, −.060], −.104 [−.148, −.064] and −.100 [−.136, −.065].

The first accepted call. At the first accepted call C_{full} and C_{last} see the same failed transcript and differ in the instruction and C_{full}'s "[Attempt 1]" label; against A, C_{full} is −.38 to −.41 and C_{last} −.27 to −.28 on Llama, and −.41 to −.49 against −.32 to −.38 on Gemma (an offline reading of Table 13); the revision-instruction policies C_{last} and H, and D, are also below A there.

Checked fragments, in full. In the nine-policy collection F shows, on its first call and over E's stand-in, up to three checked equalities from the failed attempt; its later calls are C_{last} revisions; its contrasts against A at the first accepted call and at episode end are in Tables 13 and 10, and the first-call $F - E$, the memo with the review note that carries it, is at the foot of Table 15, with no Holm rejection within family and ceiling. Whether the source of the facts matters is asked by a separate factorial on a development set: two failed draws per problem, each carrying at least one admissible equality (48 Llama and 24 Gemma problems, four episodes per start; 17.8% of Llama's and 27.8% of Gemma's failed draws carry a checkable equality). It crosses the memo — up to three checked equalities from the source attempt, as many from another problem of the same subject and level with the closest memo length (the donor rule), or none, registered as "relevant", "irrelevant" and "none" — with display: the failed attempt shown, one call at a 4k cap, or hidden under the exclusion controller and scored on the accepted draw. In every condition the memo slot follows a notice that the answer is incorrect, and the three shown memo conditions have byte-identical prompt-token totals per model. Hidden conditions add the controller, so hidden minus shown is a joint hiding-plus-exclusion effect; read at equal call opportunity (first draw only; an offline reading specified after the factorial's results) it is $+.193$ $[+.125, +.260]$, $+.146$ and $+.151$ on Llama for the three memo levels and $+.214$ $[+.083, +.349]$, $+.219$ and $+.214$ on Gemma. Source-attempt facts against no memo, with the attempt hidden, are not detected: $-.044$ $[-.094, +.005]$ on Llama and $-.052$ $[-.151, +.036]$ on Gemma; source-attempt facts are not separated from another problem's on Llama and are below them on Gemma with the attempt hidden ($-.078$ $[-.172, -.005]$).

A positive control for the memo. To ask whether the memo slot can carry content at all, the same slot (at most three equalities, 48 tokens, the same prompt and padding) was filled with intermediate results derived from the problem's reference solution. A generator model (Qwen3-32B) wrote up to three equalities the reference solution establishes before its answer; the gold guard and a string test removed any that state the answer, and a second model (Gemma-3-12B) removed any it judged not entailed, leaking the answer, or not advancing the solution (1,407 of 2,204 kept; 66 of Llama's 80 problems and 63 of Gemma's 69 reference-solved problems have at least one). A reading of the first 40 problems found the kept facts generous: four problems carry an equality one or two steps from the answer, and two of the 96 facts read there are not valid equalities. Five first-call conditions ran on the frozen starts at 4k, 16 calls per problem and condition: no memo, the source-attempt facts, an empty memo, the problem's own reference facts, and another problem's reference facts chosen by the donor rule of the factorial, the last three at one prompt length. The registered primary contrast is own-problem minus other-problem reference facts. On Llama it is $+.128$ $[+.071, +.188]$: the memo is read. The gap is mostly the other problem's facts costing $-.093$ $[-.136, -.051]$ against the empty memo; the problem's own reference facts against the empty memo are not detected, $+.035$ $[-.015, +.085]$, nor against no memo, $+.045$ $[-.006, +.094]$. On Gemma no contrast is detected (primary $+.028$ $[-.022, +.080]$; own reference facts against the empty memo $+.016$ $[-.023, +.061]$). The four differences of the no-memo and source-attempt conditions from the first calls of E and F in the nine-policy collection on the same states are not detected ($-.021$ to $+.012$). Since the problem's own reference facts are not detected above the empty memo, the non-detection for source-attempt facts in the checked-fragment reading of §5 concerns this extraction-and-check pipeline in this memo slot, not what a failed attempt contains.

Does the extractor come up empty? Stratifying the checked-fact policy by how many admissible facts a state carries, 127 of 160 Llama states and 132 of 160 Gemma states carry none. On Llama $F - A$ is below zero with the interval excluding it in every group at the 4k ceiling ($-.105/-.263/-.229/-.250$ for 0, 1, 2 and ≥ 3 facts), so an empty extractor does not explain that policy's loss; on Gemma at that ceiling every group's interval contains zero.

Table 11: The ledger, every policy but B_{budget}, over the states with $g_A > 0$ (Appendix H; Δc can therefore differ from Table 10 in the last digit). c_φ: success; g_φ: terminal answer outside the rejected class; Δc: policy minus A, split exactly into routing and proposal (bold: interval excludes zero); share: routing over Δc (shown where $|\Delta c| \geq .02$), above 1 where the proposal term has the sign opposite to Δc and negative where the routing term does; locked: cell-mean fraction of states with all eight episodes on the rejected class; ξ_φ: changed-proposal excess on the states where the policy leaves the class, with its cell count.

Family	Policy	c_φ	g_φ	Δc	routing	proposal	share	locked	ξ_φ (n)
Llama-3.1-8B	A (2k)	.56	.94	–	–	–	–	–	0
	B (2k)	.60	.99	**+.04**	**+.02**	+.01	.70	.00	+.01 (80)
	G (2k)	.54	.94	−.02	.00	−.02	.11	.00	−.02 (80)
	E (2k)	.52	.95	**−.04**	.00	**−.04**	−.02	.00	**−.05** (80)
	F (2k)	.44	.94	**−.12**	−.01	**−.12**	.07	.00	**−.12** (80)
	D (2k)	.42	.94	**−.14**	−.01	**−.14**	.05	.00	**−.15** (80)
	C_{last} (2k)	.34	.88	**−.22**	**−.04**	**−.17**	.20	.00	**−.19** (80)
	H (2k)	.34	.89	**−.22**	**−.04**	**−.18**	.17	.00	**−.21** (80)
	C_{full} (2k)	.23	.55	**−.33**	**−.23**	**−.10**	.70	.12	**−.19** (76)
	A (4k)	.57	.95	–	–	–	–	–	0
	B (4k)	.58	.99	+.01	**+.03**	−.02	–	.00	−.02 (80)
	G (4k)	.56	.94	−.01	.00	−.01	–	.00	−.01 (80)
	E (4k)	.52	.95	**−.05**	−.01	**−.05**	.13	.00	**−.05** (80)
	F (4k)	.44	.95	**−.14**	.00	**−.13**	.01	.00	**−.14** (80)
	D (4k)	.42	.95	**−.15**	.00	**−.16**	−.03	.00	**−.16** (80)
	C_{last} (4k)	.37	.89	**−.20**	**−.02**	**−.17**	.12	.01	**−.19** (80)
	H (4k)	.33	.90	**−.24**	**−.02**	**−.22**	.10	.00	**−.24** (80)
	C_{full} (4k)	.24	.56	**−.33**	**−.23**	**−.10**	.70	.12	**−.19** (74)
	A (8k)	.58	.94	–	–	–	–	–	0
	B (8k)	.60	1.00	+.02	**+.03**	.00	1.16	.00	.00 (80)
	G (8k)	.57	.94	.00	.00	.00	–	.00	.00 (80)
	E (8k)	.54	.95	**−.04**	+.01	**−.05**	−.20	.00	**−.05** (80)
	F (8k)	.42	.94	**−.16**	.00	**−.15**	.03	.00	**−.16** (80)
	D (8k)	.44	.95	**−.14**	+.01	**−.14**	−.04	.00	**−.15** (80)
	C_{last} (8k)	.39	.92	**−.19**	−.01	**−.19**	.04	.00	**−.20** (80)
	H (8k)	.40	.92	**−.18**	−.01	**−.16**	.06	.00	**−.18** (80)
	C_{full} (8k)	.24	.54	**−.34**	**−.24**	**−.10**	.70	.14	**−.17** (77)
Gemma-3-12B	A (2k)	.58	.88	–	–	–	–	–	0
	B (2k)	.62	.92	**+.04**	**+.03**	+.01	.82	.00	+.01 (80)
	G (2k)	.58	.89	.00	+.01	−.01	–	.01	−.01 (80)
	E (2k)	.55	.87	−.02	−.01	−.02	.24	.01	−.01 (80)
	F (2k)	.57	.89	−.01	+.01	−.01	–	.00	.00 (80)
	D (2k)	.49	.72	**−.09**	**−.08**	−.01	.92	.01	+.01 (80)
	C_{last} (2k)	.33	.53	**−.25**	**−.23**	−.02	.92	.09	−.01 (77)
	H (2k)	.38	.60	**−.20**	**−.18**	−.02	.90	.05	−.02 (77)
	C_{full} (2k)	.18	.29	**−.39**	**−.40**	+.01	1.02	.35	−.01 (60)
	A (4k)	.63	.88	–	–	–	–	–	0
	B (4k)	.68	.96	**+.05**	**+.05**	.00	1.02	.00	.00 (80)
	G (4k)	.63	.88	.00	.00	.00	–	.02	+.01 (79)
	E (4k)	.59	.88	**−.05**	−.01	**−.04**	.17	.00	**−.05** (80)
	F (4k)	.64	.90	.00	+.02	−.01	–	.00	−.01 (80)
	D (4k)	.59	.85	**−.04**	−.02	−.02	.52	.01	−.03 (80)
	C_{last} (4k)	.48	.69	**−.16**	**−.13**	−.02	.86	.04	−.03 (79)
	H (4k)	.51	.74	**−.12**	**−.10**	−.02	.86	.02	−.05 (80)
	C_{full} (4k)	.24	.39	**−.39**	**−.37**	−.02	.95	.26	**−.09** (65)
	A (8k)	.62	.87	–	–	–	–	–	0
	B (8k)	.69	.99	**+.07**	**+.08**	.00	1.03	.00	.00 (80)
	G (8k)	.62	.88	.00	+.01	−.01	–	.00	−.01 (80)
	E (8k)	.60	.88	−.02	.00	−.02	−.21	.00	−.03 (80)
	F (8k)	.65	.90	+.03	+.02	+.01	.56	.00	+.02 (80)
	D (8k)	.62	.87	+.01	+.01	.00	–	.01	−.01 (80)
	C_{last} (8k)	.56	.81	−.05	**−.05**	.00	.96	.02	−.02 (80)
	H (8k)	.58	.82	−.03	**−.04**	.00	1.10	.01	−.02 (80)
	C_{full} (8k)	.28	.43	**−.34**	**−.34**	.00	1.01	.28	−.03 (64)
Qwen3-32B	A (2k)	.64	.94	–	–	–	–	–	0
	F (2k)	.64	.93	.00	−.01	+.01	–	.01	+.01 (80)
	C_{full} (2k)	.18	.28	**−.46**	**−.42**	**−.04**	.91	.46	**−.16** (58)
	A (4k)	.69	.92	–	–	–	–	–	0
	B (4k)	.72	.98	**+.04**	**+.04**	.00	1.14	.00	.00 (80)
	F (4k)	.71	.94	+.02	+.01	+.01	–	.00	+.01 (80)
	C_{full} (4k)	.21	.30	**−.47**	**−.44**	**−.03**	.94	.44	**−.17** (58)
	A (8k)	.70	.92	–	–	–	–	–	0
	F (8k)	.72	.95	+.02	**+.02**	.00	–	.00	.00 (80)
	C_{full} (8k)	.23	.31	**−.48**	**−.44**	**−.04**	.92	.40	**−.16** (60)

Table 12: **The whole history against the verdict-only restart B.** The ledger of Table 11 paired with B on the same states: $c_\varphi - c_B = (g_\varphi - g_B)v_y + (P_\varphi - P_B)$, exact state by state, with B's measured proposal term. 80 cells per row; paired cell bootstrap 95% intervals. Offline analysis, specified after the ledger was read. B stops at its first accepted draw and C_{full} runs to the ceiling, so these remain whole-policy contrasts.

C_{full}, family and ceiling	$c_\varphi - c_B$	routing against B	proposal against B
Llama-3.1-8B, 2k	$-.370$ $[-.436, -.306]$	$-.257$ $[-.316, -.204]$	$-.112$ $[-.156, -.072]$
Llama-3.1-8B, 4k	$-.341$ $[-.408, -.271]$	$-.262$ $[-.325, -.203]$	$-.079$ $[-.122, -.034]$
Llama-3.1-8B, 8k	$-.366$ $[-.432, -.304]$	$-.267$ $[-.321, -.217]$	$-.098$ $[-.139, -.059]$
Gemma-3-12B, 2k	$-.434$ $[-.520, -.351]$	$-.434$ $[-.509, -.364]$	$+.001$ $[-.037, +.038]$
Gemma-3-12B, 4k	$-.439$ $[-.526, -.353]$	$-.421$ $[-.488, -.351]$	$-.018$ $[-.059, +.021]$
Gemma-3-12B, 8k	$-.411$ $[-.499, -.325]$	$-.416$ $[-.490, -.343]$	$+.005$ $[-.036, +.046]$
Qwen3-32B, 4k	$-.510$ $[-.584, -.435]$	$-.486$ $[-.557, -.418]$	$-.024$ $[-.057, +.008]$

Table 13: First accepted call against episode end, three ceilings; chain: end minus first accepted call. Cell bootstrap; bold: interval excludes zero.

Policy	Llama-3.1-8B			Gemma-3-12B			Qwen3-32B		
	first $- A$	end $- A$	chain	first $- A$	end $- A$	chain	first $- A$	end $- A$	chain
ceiling 2k									
B	**+.04**	**+.04**	.00	**+.04**	**+.04**	.00	–	–	–
G	.00	−.02	−.02	−.01	.00	+.01	–	–	–
B_{budget}	**+.03**	.00	**−.03**	+.02	+.02	.00	–	–	–
E	−.02	**−.04**	−.02	+.01	−.02	**−.03**	–	–	–
F	**−.05**	**−.12**	**−.08**	−.02	−.01	**+.01**	+.01	+.01	.00
D	**−.13**	**−.14**	−.02	**−.20**	**−.09**	**+.11**	–	–	–
C_{last}	**−.27**	**−.22**	**+.05**	**−.32**	**−.25**	**+.07**	–	–	–
H	**−.28**	**−.22**	**+.06**	**−.24**	**−.20**	**+.05**	–	–	–
C_{full}	**−.38**	**−.33**	**+.05**	**−.41**	**−.39**	**+.01**	**−.56**	**−.46**	**+.10**
ceiling 4k									
B	+.01	+.01	.00	**+.06**	**+.06**	.00	**+.04**	**+.04**	.00
G	−.02	−.01	+.01	.00	+.01	+.01	–	–	–
B_{budget}	.00	+.01	.00	**+.04**	**+.04**	.00	–	–	–
E	−.03	**−.05**	−.03	.00	**−.04**	**−.05**	–	–	–
F	**−.07**	**−.14**	**−.07**	**−.04**	.00	**+.05**	−.01	+.02	**+.03**
D	**−.12**	**−.15**	−.03	**−.25**	**−.04**	**+.21**	–	–	–
C_{last}	**−.28**	**−.20**	**+.08**	**−.38**	**−.15**	**+.23**	–	–	–
H	**−.29**	**−.24**	**+.05**	**−.28**	**−.12**	**+.16**	–	–	–
C_{full}	**−.40**	**−.33**	**+.06**	**−.49**	**−.39**	**+.10**	**−.61**	**−.47**	**+.13**
ceiling 8k									
B	+.02	+.02	.00	**+.07**	**+.07**	.00	–	–	–
G	−.02	.00	+.01	+.01	.00	−.01	–	–	–
B_{budget}	+.01	+.03	+.02	**+.09**	**+.08**	−.01	–	–	–
E	−.01	**−.04**	**−.04**	**+.04**	−.02	**−.06**	–	–	–
F	**−.06**	**−.16**	**−.10**	**−.04**	+.03	**+.07**	−.01	+.02	**+.03**
D	**−.10**	**−.14**	−.03	**−.23**	+.01	**+.24**	–	–	–
C_{last}	**−.28**	**−.19**	**+.08**	**−.35**	−.05	**+.30**	–	–	–
H	**−.28**	**−.18**	**+.10**	**−.26**	−.03	**+.22**	–	–	–
C_{full}	**−.41**	**−.34**	**+.07**	**−.48**	**−.34**	**+.14**	**−.62**	**−.48**	**+.14**

Table 14: **The paired effort contrasts on gpt-oss-120b.** Two problem sets, the low-effort set (80 cells, upper block) and the high-effort set (61 cells, lower block); on each, the three endpoint policies at reasoning effort low and high on the same cells, 32k ceiling. Row A: fresh-draw success at each effort and its paired difference. Rows B and C_{full}: policy minus A within each effort, and the difference of differences (high minus low). Seeded cell bootstrap 95% intervals; bold: interval excludes zero. Truncated episodes count as failures; C_{full} at low effort reaches the 64-call cap in 71.8% (upper) and 65.1% (lower) of episodes, counted by the final answer here and as failures in the last column of Table 4.

Policy	keeps	effort low	effort high	high − low
low-effort set, 80 cells				
A	nothing (one fresh draw)	.505	.884	**+.378 [+.317, +.443]**
B	exclusion of y's class only	**+.084 [+.054, +.116]**	+.012 [−.004, +.030]	**−.073 [−.105, −.041]**
C_{full}	whole answer history	**−.300 [−.370, −.229]**	**−.076 [−.116, −.040]**	**+.224 [+.155, +.298]**
high-effort set, 61 cells				
A	nothing (one fresh draw)	.217	.612	**+.394 [+.322, +.463]**
B	exclusion of y's class only	−.006 [−.036, +.025]	+.016 [−.017, +.048]	+.023 [−.020, +.064]
C_{full}	whole answer history	**+.091 [+.004, +.185]**	**−.101 [−.175, −.032]**	**−.193 [−.261, −.117]**

Table 15: **Content × display.** Cell-mean success by condition and the registered contrasts (paired cells, seeded cell bootstrap; bold: interval excludes zero), with the held-out first-call 2 × 2 on the nine-policy episodes, whose content $F - E$ is the memo with the review note that carries it.

	Llama-3.1-8B	Gemma-3-12B
cells (two starts × four episodes)	48	24
no memo, shown	.167	.245
source-attempt facts, shown	.164	.161
other-problem facts, shown	.161	.219
no memo, hidden	.372	.516
source-attempt facts, hidden	.328	.464
other-problem facts, hidden	.326	.542
reference: whole history (C_{full})	.044	.078
reference: content-free stand-in (E), hidden	.391	.505
reference: judged prefix (D), judge failures included	.302	.281
source-attempt − other-problem, shown	+.003 [−.047, +.052]	−.057 [−.151, +.026]
source-attempt − other-problem, hidden	+.003 [−.068, +.073]	**−.078 [−.172, −.005]**
hidden − shown, no memo	**+.206 [+.143, +.273]**	**+.271 [+.151, +.401]**
hidden − shown, source-attempt facts	**+.164 [+.096, +.234]**	**+.302 [+.167, +.443]**
hidden − shown, other-problem facts	**+.164 [+.091, +.245]**	**+.323 [+.193, +.464]**
C_{full} − no memo, shown	**−.122 [−.180, −.070]**	**−.167 [−.234, −.094]**
E − no memo, hidden	+.018 [−.029, +.062]	−.010 [−.094, +.068]
D − no memo, shown	**+.135 [+.070, +.201]**	+.036 [−.047, +.115]
interaction (src−oth)$_{\text{hidden}}$−(src−oth)$_{\text{shown}}$	+.000 [−.083, +.094]	−.021 [−.109, +.057]
Held-out 2 × 2 at the first call, before any redraw (nine-policy episodes; Holm within family and ceiling)		
content $F - E$, 2k	+.000 [−.033, +.032]	−.010 [−.041, +.021]
display $E - C_{\text{last}}$, 2k	**+.223 [+.160, +.283]**	**+.312 [+.238, +.382]**
content $F - E$, 4k	−.027 [−.061, +.007]	+.007 [−.023, +.038]
display $E - C_{\text{last}}$, 4k	**+.236 [+.181, +.291]**	**+.329 [+.261, +.405]**
content $F - E$, 8k	−.016 [−.050, +.015]	−.007 [−.038, +.024]
display $E - C_{\text{last}}$, 8k	**+.245 [+.192, +.302]**	**+.320 [+.249, +.390]**

I BUDGET, TRUNCATION AND MULTIPLICITY

Truncation. The main rule grades an episode's last completed call (B_{budget}: its last accepted draw) and scores an episode with none as a failure, the convention the protocol froze. Three readings are computed for A and each retry policy but B_{budget} at every ceiling on Llama-3.1-8B and Gemma-3-12B, the first two also for B_{budget} and Qwen3-32B: the main rule, the analysis's cap-hit-as-failure column, and an adversarial recount in which every episode with no completed call is scored correct. Under the adversarial recount the policy ordering is unchanged in four of the six (family, ceiling) groups, and the whole history's gap against A falls to half its main-rule value in 0 of those six groups. Gemma-3-12B's C_{full} has no completed call in 1.3%, 0.0% and 0.1% of episodes at the three ceilings. Cap-hit readings of the confirmation regimes: Table 4.

Multiplicity. A first attempt at a bootstrap-t max-T adjustment degenerated on $0/1$ data, since individual resamples give a near-zero standard error and an adjusted critical value of 14.02, rejecting nothing. The replacement, registered after the primary results had been read (Appendix G), is a single-step Westfall–Young max-T with fixed standardization: the resample statistic is $t^*_{\ell,k} = (\bar{d}^*_{\ell,k} - \bar{d}_k)/(s_k/\sqrt{n_k})$, with $\bar{d}_k$ the mean paired difference of contrast k over its n_k cells, ℓ the resample index and s_k the original sample's standard deviation; the threshold is the 95th percentile of $\max_k |t^*_{\ell,k}|$ over one joint resample stream, and the observed $|t_k| = |\bar{d}_k|/(s_k/\sqrt{n_k})$ is compared against it. The two families are resampled at shared cell positions, each on its own 80 confirmation problems; an earlier version mapped Gemma's problems onto Llama-3.1-8B's identifiers and diluted Gemma's differences, and was discarded after its output had been read. The threshold is 3.18, 32 of the 48 contrasts against A are rejected unadjusted and 24 after adjustment (marked in Table 10). The analysis output carries 54 rows because it includes the six $A - A$ rows, which are zero by construction and do not enter the maximum. An offline check specified after the adjusted verdicts were read draws the index vector independently for each family; the threshold moves to 3.20 and every adjusted verdict is unchanged. B_{budget}, collected after the adjustment was run, is not in the adjustment set; its bold marks in Table 10 rest on unadjusted intervals.

J THE SELF-CORRECTION TRAINING FRONTIER

Design. S2R (Ma et al., 2025) publishes a process-RL (PRL) and an outcome-RL (ORL) checkpoint over the same Llama base and chat template, which permits the one comparison the ledger needs: each checkpoint is read against *its own* clean-context law, since training changes p_0 and therefore changes v_y. Two readings are reported. The *pooled* reading, offline, uses archived collections: per cell, 64 clean-context samples and sixteen verdict-free revision chains at 16k (twelve of 8 steps, four of 16) from each checkpoint's own wrong answers, 105 cells per checkpoint. The *common-start* reading runs all three checkpoints from the same frozen wrong starts, two on each of 80 problems, under the S2R study's exclusion controller (Appendix E). The checkpoints are used as revisers, not under S2R's own evaluation: a revision is one call that sees the frozen start, a wrong answer of the untrained base, with the revision instruction and a notice that the answer is incorrect, sampled at temperature .8 (top-p .95) under a 4,096-token cap (16,384 in the budget check), whereas Ma et al. (2025) evaluate greedy verify-then-correct trajectories with no external verdict. The notice supplies the verdict that S2R trains the model to reach itself.

The pooled margin. On the archived verdict-free chains the per-cell margin M_y averages, over steps whose previous answer is wrong, success minus the checkpoint's own verdict-only reference (the filtered-pool form of $\hat{q}_0/(1 - \hat{p}_0(y))$): $-.362$ $[-.393, -.329]$, $-.276$ $[-.313, -.239]$ and $-.185$ $[-.217, -.153]$ for the base, PRL and ORL, with paired increments $+.086$ $[+.041, +.130]$ (PRL−base) and $+.177$ $[+.136, +.217]$ (ORL−base). On the same chains the classical turn-2 minus turn-1 statistic of Kumar et al. (2025), which also scores revisions of answers that were already correct and sees no verdict, puts the three checkpoints in the other order: $-.144$ $[-.174, -.115]$, $-.187$ $[-.231, -.145]$, $-.196$ $[-.246, -.148]$, with increments $-.043$ $[-.095, +.008]$ and $-.052$ $[-.106, +.001]$, neither detected. A change of ranking between the two was the collection's registered criterion. The correction rate of §6, one-step revision from the base's frozen starts with the verdict supplied, is a different object from both.

On common starts. Table 16 gives the three checkpoints on the shared starts of the 80 problems, with the sign convention of Figure 4: revision stays below the checkpoint's own verdict-only restart

at all three checkpoints, every interval excluding zero. At the two trained checkpoints the restart is not separated from the fresh draw ($+.018$ $[-.009, +.048]$ for PRL, $+.023$ $[-.009, +.055]$ for ORL; restart minus fresh draw). Between the two trained checkpoints, revision is not separated ($+.005$ $[-.034, +.048]$ for ORL$-$PRL); they are two recipes on one 8B base, not two points of a training curve, and the adversarial truncation reading, in which every unfinished episode is scored correct, puts the process-RL checkpoint's revision ahead because it truncates more often.

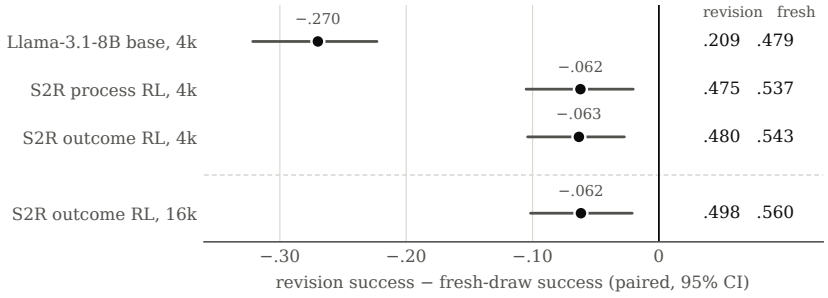

Figure 4: **Revision success rises; the revised retry stays below a fresh draw from the same checkpoint.** Revision success minus one unconstrained fresh draw from the same checkpoint's clean context, on common frozen wrong starts of 80 problems, with paired problem-level bootstrap 95% intervals; right: the two success rates. The 16k row repeats the outcome-RL checkpoint with a 16k per-call cap, the fresh draw collected rather than simulated.

Table 16: **The base and S2R's two RL checkpoints, on common starts.** One-step revision, a fresh draw and the verdict-only restart on the same frozen wrong starts (80 problems, two each), then revision minus that restart within each checkpoint (Figure 4 plots revision minus the fresh draw), with cell bootstrap 95% intervals. The margin column M_y is the pooled offline reading over 105 cells: the success of a verdict-free revision step from the checkpoint's own wrong answer minus that checkpoint's own verdict-only reference.

Checkpoint	revision	fresh draw	restart	revision − restart	margin M_y
base	.209	.479	.484	$-.275$ $[-.339, -.214]$	$-.362$ $[-.393, -.329]$
S2R-PRL	.475	.537	.555	$-.080$ $[-.123, -.036]$	$-.276$ $[-.313, -.239]$
S2R-ORL	.480	.543	.566	$-.086$ $[-.136, -.039]$	$-.185$ $[-.217, -.153]$

Would a longer cap close the gap? At 4k, ORL's revision truncates in .191 of common-start episodes. An offline analysis registered before the collection that tested it put the rescue rate on truncated episodes needed to close the gap at .70 $[.38, 1.11]$ for ORL and .41 for PRL, against fresh-draw pool rates of .22 and .14 (above one at base), and predicted a 16k revision success of .502 and a gap of $-.090$ against the verdict-only restart. The collection at a 16k per-call cap on ORL, with a same-node 4k control ($+.002$ $[-.050, +.050]$ against the earlier 4k run), cut truncation from .205, the control's rate, to .011 and moved revision success by $+.017$ $[-.028, +.064]$, to .498 (net rescue .10 $[-.18, +.38]$). The verdict-only restart and fresh draw, collected at 16k on the same nodes, are .584 and .560, within .005 of the pool simulation (intervals containing zero). Revision stays below both: $-.086$ $[-.130, -.041]$ against the restart and $-.062$ $[-.102, -.021]$ against the fresh draw, the values in the main text and Figure 4.

K THE SHARED-CONTROLLER STUDY

Design. The study measures the retention contrast of §3: D_{ret} keeps the whole transcript and $D_{\varnothing}$, the withheld condition, holds the one-line withholding notice in its place, both under the same exclusion controller. On Llama-3.1-8B it runs 161 problems at an 8k ceiling, two frozen wrong starts and eight episodes each; the controller allows 32 calls with no forced acceptance and scores truncation as terminal failure. The study was collected as a transcript $\times$ hint design, with one gold-derived fact as the hint, whose registered target was the transcript $\times$ hint interaction (Appendix G);

J on Llama-3.1-8B is therefore a discovery inside that design, one of the five contrasts its frozen readout reports, each with a Bonferroni-corrected interval at level $1 - .05/6$ (the readout's six-test family), not the target quantity. All four conditions are in Table 17. The 161 problems merge a 64-problem development set with 97 added afterwards: the rest of the eligible problems in the same screened set, once 17 used in an earlier pilot were set aside. On the development set the hint check was evaluated on the two withheld conditions, and J was read on a 16- and then a 32-problem probe from which the powered collection was sized; the 97 were collected only after that.

Result. Cell means are .568 for $D_{\varnothing}$ and .333 for D_{ret}; $D_{\text{ret}} - D_{\varnothing}$ is $-.235$ $[-.276, -.194]$ on the merged 161 problems and $-.210$ $[-.261, -.161]$ on the 97 added afterwards, a split made after the merged results (problem-paired bootstrap, 95%). With the hint in both conditions J is $-.241$ $[-.283, -.202]$; the hint raises the withheld retry by $+.033$ $[+.008, +.059]$ and the kept one by $+.027$ $[-.006, +.061]$, and their difference, the registered interaction, is not detected ($-.007$ $[-.046, +.034]$; corrected $[-.060, +.050]$). On the added 97 the hint gain on the withheld retry is not detected ($+.017$ $[-.013, +.047]$). Truncation is .055 on $D_{\varnothing}$ and .115 on D_{ret} at the 8k ceiling.

Table 17: **The four conditions of the Llama-3.1-8B collection** (8k, 16 episodes per problem and condition). Transcript kept (D_{ret}) or withheld ($D_{\varnothing}$), crossed with one gold-derived fact as a hint or none; problem-paired bootstrap 95% intervals, and for the interaction also the Bonferroni-corrected interval (level $1 - .05/6$, six-test family). The registered target was the interaction; the 97-problem column is a split made after the merged results.

	merged 161	added 97
success, withheld, no hint ($D_{\varnothing}$)	.568	.524
success, withheld, hint	.602	.541
success, kept, no hint (D_{ret})	.333	.314
success, kept, hint	.360	.304
J, no hint (kept − withheld)	$-.235$ $[-.276, -.194]$	$-.210$ $[-.261, -.161]$
J, hint	$-.241$ $[-.283, -.202]$	$-.237$ $[-.290, -.185]$
hint gain, withheld	$+.033$ $[+.008, +.059]$	$+.017$ $[-.013, +.047]$
hint gain, kept	$+.027$ $[-.006, +.061]$	$-.010$ $[-.048, +.031]$
interaction (hint gain, kept − withheld)	$-.007$ $[-.046, +.034]$; corrected $[-.060, +.050]$	$-.027$ $[-.070, +.018]$

Further models. The two no-hint conditions were then run, with unchanged rules, on Gemma-3-12B, Qwen3-8B and Qwen3-32B (thinking off), as prospective replications: each registered J as its primary contrast, with the problem set and decision rule fixed before any episode. Each model was first screened on the same 700-problem candidate set, 64 draws per problem; a problem is eligible when that pool holds at least two complete wrong draws, two of which are frozen as its starts. Gemma-3-12B uses the 85 eligible problems that no earlier Gemma collection had touched; for the Qwen3 models the first 100 eligible problems in candidate order form the set (234 and 276 were eligible) and the next 8 a truncation pilot. The registered criterion was a truncated-final rate below .10 on the pilot at 8k; Gemma-3-12B passed, and Qwen3-8B and Qwen3-32B failed and ran both policies at a 16k ceiling (Appendix G). Cell means are .575 for $D_{\varnothing}$ and .451 for D_{ret} on Gemma-3-12B ($-.124$ $[-.179, -.071]$), .697 and .577 on Qwen3-8B ($-.120$ $[-.175, -.065]$), and .735 and .629 on Qwen3-32B ($-.106$ $[-.157, -.058]$). Truncation on $D_{\varnothing}$ and D_{ret} is .062 and .212 for Gemma-3-12B at 8k, .035 and .155 for Qwen3-8B and .031 and .091 for Qwen3-32B at 16k. Because the Gemma-3-12B main run truncated more than its pilot had, both of its policies were also collected at 16k on the same problems: truncation falls to .018 and .090 and $D_{\text{ret}} - D_{\varnothing}$ is $-.111$ $[-.168, -.054]$ (Table 21). In all three models D_{ret} spends about twice the output tokens of $D_{\varnothing}$ per episode (3.7k–5.8k against 1.8k–2.5k).

What enforcing the rejection recovers. Under this controller (Appendix E) a first call is redrawn only if it falls in the rejected class, so $D_{\text{ret}} - D_{\varnothing}$ splits into the first call accepted and correct, what enforcing the rejection adds, and the terminal failure states (Table 2); the split is an offline analysis specified after the contrasts above had been read. With the transcript in context the first call returns to the rejected class far more often than $D_{\varnothing}$'s first call (.20 against .03 on Llama-3.1-8B, .61 against .14 on Gemma-3-12B, .67 against .17 on Qwen3-8B, .53 against .14 on Qwen3-32B; differences $+.173$ $[+.149, +.199]$, $+.462$ $[+.403, +.521]$, $+.497$ $[+.444, +.548]$ and $+.391$ $[+.342, +.439]$, problem-paired bootstrap 95%).

Where the loss lands. On Gemma-3-12B and Qwen3-8B the loss sits in the budget states, since each return to the rejected class is a billed call, and the difference in accepted wrong answers has an interval containing zero; on Qwen3-32B the budget states carry .064 of the .106. Doubling Gemma-3-12B's ceiling lengthens the loop instead of ending it: rejected calls per episode rise from 3.6 to 5.2, and part of the loss moves to the 32-call cap (Table 21). On Llama-3.1-8B, .132 of the .235 is an accepted answer that is new and wrong, and its budget states carry .023. Its first call that leaves the rejected class is also correct less often than $D_\varnothing$'s: $-.224$ $[-.269, -.180]$ on Llama-3.1-8B, with intervals containing zero on the other three models ($-.026$ $[-.107, +.054]$ on Gemma-3-12B, $-.008$ $[-.061, +.042]$ on Qwen3-8B, $-.020$ $[-.072, +.028]$ on Qwen3-32B; registered reading, episodes pooled); a cell-weighted reading added after the first run gives $-.215$ $[-.260, -.172]$ on Llama-3.1-8B and $-.051$ $[-.098, -.006]$ on Qwen3-32B, the ordering of the proposal terms in Table 12.

What is shown: the answer and the reasoning. Five further conditions on Llama-3.1-8B at 8k, under the same controller, revision instruction and exclusion in two registrations, ask which part of the transcript carries the loss. They run on the 130 of the 161 problems whose two starts both carry a boxed answer (Appendix G), with $D_\varnothing$ and D_{ret} re-collected the same day (Appendix E): $D_{\text{ret}} - D_\varnothing$ is $-.237$ $[-.280, -.196]$ against $-.226$ $[-.273, -.182]$ for the original collection on the same problems, the registered comparison. The conditions are in Table 18. The first registration's primary contrast, D_{ret} minus the answer line alone, is not detected ($+.041$ $[-.001, +.081]$). That condition ends .314 of its episodes unparseable, so it is not a clean removal of the reasoning (Appendix G); against the answer line after a content-free placeholder of matched token length, D_{ret} is $-.100$ $[-.146, -.058]$. The second completes the 2×2 with strict redaction (every boxed expression, and the answer string in the last three non-empty lines); a third condition redacts boxes only. Primary, redacted transcript minus placeholder: $-.260$ $[-.303, -.215]$. Transcript minus redacted: $-.017$ $[-.046, +.012]$. The two redactions: $-.005$ $[-.030, +.020]$. The answer on the placeholder: $-.177$ $[-.217, -.138]$. The factors are sub-additive (interaction $+.160$ $[+.112, +.211]$; main effects $-.180$ $[-.218, -.144]$ for the reasoning and $-.097$ $[-.122, -.072]$ for the answer), also in the first call's return (interaction $-.215$ $[-.256, -.174]$). The answer string survives strict redaction somewhere in 36.5% of transcripts. On the 103 problems with states where it occurs nowhere, the redacted transcript's first call returns at .211 against .025 ($+.185$ $[+.148, +.226]$; added after the first results), and its success is $-.255$ $[-.311, -.201]$ against the placeholder. A first call that leaves the class is correct less often with the reasoning ($-.230$ $[-.280, -.183]$ with the answer redacted, $-.135$ $[-.186, -.086]$ with it visible) and, by less, with the answer ($-.083$ $[-.125, -.040]$ on the placeholder; $+.011$ $[-.031, +.052]$ with the reasoning shown). Correct but unboxed finals in the redacted conditions (.000 and .001 of episodes) make a lenient analysis equal the registered one to .001, and no other mathematics condition has an unparseable final that the grader scores correct.

Table 18: **What the retained turn shows** (Llama-3.1-8B, 8k, 130 problems, 16 episodes per problem and condition). The four middle rows form the 2×2 at one instruction and a matched prompt length. First call leaves: correct among first calls that leave the rejected class, problem-weighted.

Shown	success	first call returns	first call leaves: correct	accepted wrong	unparseable	full ceiling
withholding notice ($D_\varnothing$)	.557	.027	.614	.294	.080	.069
placeholder	.596	.027	.624	.321	.033	.050
placeholder, answer line	.419	.225	.544	.241	.135	.148
transcript, answer redacted	.337	.230	.395	.486	.090	.070
transcript (D_{ret})	.320	.212	.409	.448	.102	.107
answer line alone	.278	.212	.514	.175	.314	.177
transcript, boxed expressions redacted	.332	.187	.392	.469	.097	.089

What is shown: three more models. A third registration runs the 2×2 alone (the four middle rows of Table 18, same collector, controller, instruction and strict redaction) on Gemma-3-12B at 8k and on Qwen3-8B and Qwen3-32B at 16k with thinking off, on the frozen starts of the collections of Table 2 whose two starts both carry a boxed answer: 85, 100 and 99 problems, 16 episodes per problem and condition. One Qwen3-32B problem whose start nests one box in another, which the redaction cannot clear, was removed from all four conditions (Appendix G). The primary contrast, the two secondary contrasts, the main effects and interaction, the first-call and terminal-state analyses and the split by residual exposure were registered before collection, per model, with no pooling across models. D_{ret} was collected again; its difference from the original collection is not detected

in any model ($+.004$ $[-.019, +.028]$, $+.002$ $[-.017, +.020]$, $+.012$ $[-.009, +.034]$; hardware in Appendix E). Unparseable and full-ceiling endings are at most .013 and .009 in any condition.

Table 19 gives the conditions and the three contrasts. The answer on the placeholder has a detected cost on Llama only; its interval contains zero in the other three, as does the interaction's ($-.009$ $[-.058, +.040]$, $+.004$ $[-.034, +.042]$, $+.030$ $[-.006, +.066]$), where on Llama the two factors are sub-additive. The first call's return to the rejected class is raised by either factor in every model, by the reasoning more than by the answer ($+.313$, $+.416$, $+.334$ against $+.238$, $+.177$, $+.151$, all intervals excluding zero), with a negative interaction throughout ($-.126$ $[-.185, -.068]$, $-.077$ $[-.128, -.030]$, $-.095$ $[-.137, -.055]$). After strict redaction the answer string still occurs somewhere in 61.8%, 58.0% and 57.1% of the transcripts. On the 45, 52 and 57 problems with states where it occurs nowhere, the redacted transcript's first call returns to the rejected class at .388, .502 and .413 against .146, .186 and .121 for the placeholder ($+.242$ $[+.160, +.326]$, $+.316$ $[+.222, +.411]$, $+.293$ $[+.221, +.364]$); its success on those problems is $-.060$ $[-.114, -.006]$, $-.067$ $[-.143, +.002]$ and $-.138$ $[-.224, -.053]$ against the placeholder.

Where the reasoning's loss lands differs by model, as for the whole transcript in Table 2. With the answer redacted, the reasoning raises the budget states by $+.093$ $[+.055, +.131]$ on Gemma-3-12B and $+.089$ $[+.046, +.136]$ on Qwen3-8B, where its difference in accepted wrong answers has an interval containing zero ($+.021$ $[-.030, +.074]$, $-.016$ $[-.059, +.026]$); on Qwen3-32B it raises both ($+.035$ $[+.009, +.062]$ and $+.072$ $[+.033, +.112]$); on Llama-3.1-8B the loss is an accepted wrong answer ($+.165$ $[+.118, +.211]$ against $+.017$ $[+.010, +.025]$ in the budget states). A first call that leaves the class is correct less often with the reasoning by $-.230$ on Llama-3.1-8B, $-.081$ $[-.140, -.025]$ on Qwen3-32B and $-.059$ $[-.115, -.004]$ on Gemma-3-12B, and not detectably on Qwen3-8B ($-.023$ $[-.077, +.031]$). On Gemma-3-12B the answer line moves episodes between failure states without a detected change in success: accepted wrong answers fall ($-.143$ $[-.190, -.095]$ on the placeholder) and the budget states rise ($+.130$ $[+.080, +.179]$).

Four offline checks were specified after the results. (i) Success recomputed from the raw records by a script that shares no code with the analysis (an episode succeeds if it holds a call that is accepted and correct) reproduces all sixteen (model, condition) success rates exactly; every episode has one final call, none accepts an answer inside the rejected class, and none exceeds its ceiling. (ii) Against each model's withheld condition $D_{\varnothing}$ (the same-day collection on Llama-3.1-8B, the original collection elsewhere), whose instruction differs from the one the placeholder and transcript conditions share, the placeholder is $+.039$ $[+.016, +.062]$, $-.026$ $[-.060, +.008]$, $-.016$ $[-.042, +.009]$ and $+.008$ $[-.015, +.032]$, and the redacted transcript is $-.220$ $[-.264, -.177]$ (registered for Llama-3.1-8B), $-.135$ $[-.188, -.087]$, $-.090$ $[-.147, -.033]$ and $-.101$ $[-.150, -.055]$ against it. (iii) At the Bonferroni level for four primary contrasts (98.75%, 20,000 resamples) the intervals are $[-.315, -.204]$, $[-.176, -.044]$, $[-.137, -.012]$ and $[-.169, -.054]$. (iv) Prompt lengths: the two placeholder conditions are within 3 tokens of the transcript condition on Llama-3.1-8B and Gemma-3-12B and 17 to 21 tokens longer on average on the two Qwen3 models (about 1% of the prompt; at most 165), and the redacted transcript differs from the transcript by under 3 tokens on average (-136 to $+46$). Matching the answer string after stripping LaTeX spacing and fraction variants moves the no-residual subsets by at most two problems and changes none of their readings.

Another problem's transcript. A fourth registration fills the retained turn with the strictly redacted failed transcript of *another* problem from the same model's list, the one nearest in redacted token count. Ties go by problem identifier, and a donor whose answer would be graded correct on the recipient, or falls in its rejected class, is skipped. Everything else, including the recipient's rejected class, is as in Table 19. The placeholder was re-collected the same day, with token-identical prompts; it is within .009 of the original in every model. The primary contrast is the own redacted transcript (original collection) minus the other's; the secondary is the other's minus the same-day placeholder (Table 20). Donors differ from recipients by .009 to .014 of the recipient's tokens on average (95th percentile at most .040). Dropping the two to four states per model beyond 10% moves no contrast by more than .004 (a check added after the first results). In no model does the other problem's transcript raise the first call's return detectably above the placeholder's (.033, .151, .052, .100 against .029, .153, .161, .134; Llama-3.1-8B $+.004$ $[-.007, +.015]$): the return is specific to the own transcript. On Llama-3.1-8B the own transcript's extra cost runs through a first call that leaves the class and is correct less often ($-.188$ $[-.232, -.141]$) and through more accepted wrong answers ($+.136$ $[+.094, +.176]$). The other transcript's cost runs by the same route

Table 19: **What the retained turn shows, four models.** Success in the four conditions of the 2×2 (one instruction, matched prompt length), the three contrasts with problem-paired bootstrap 95% intervals, and the share of first calls that return to the rejected class in the same four conditions; n: problems, 16 episodes per problem and condition. Llama-3.1-8B is the second registration of Table 18.

Model	n	placeholder		transcript		reasoning alone (redacted − placeholder)	answer added (full − redacted)	answer alone (answer − placeholder)
		no answer	answer	redacted	full			
Llama-3.1-8B, 8k	130	.596	.419	.337	.320	−.260 [−.303, −.215]	−.017 [−.046, +.012]	−.177 [−.217, −.138]
Gemma-3-12B, 8k	85	.549	.574	.440	.455	−.110 [−.163, −.060]	+.015 [−.022, +.051]	+.024 [−.011, +.061]
Qwen3-8B, 16k	100	.681	.648	.607	.579	−.074 [−.124, −.027]	−.028 [−.068, +.010]	−.032 [−.071, +.001]
Qwen3-32B, 16k	99	.751	.727	.641	.648	−.109 [−.158, −.064]	+.006 [−.027, +.037]	−.023 [−.064, +.012]
First call returns to the rejected class								
Llama-3.1-8B, 8k		.027	.225	.230	.212			
Gemma-3-12B, 8k		.156	.394	.469	.581			
Qwen3-8B, 16k		.161	.338	.576	.676			
Qwen3-32B, 16k		.126	.277	.460	.516			

(−.067 [−.094, −.041]; +.051 [+.025, +.080]) and is .185 [.084, .303] of the redacted transcript's loss. On the other three models the registered reading is not applied (Appendix G): a diagnostic written after the first two results found that those retries answer the donor's problem (last column of Table 20; chance at most .002). In .051, .152 and .421 of episodes on Gemma-3-12B, Qwen3-32B and Qwen3-8B the accepted answer is the donor's own redacted answer, regenerated (first call .034, .138, .401; Llama-3.1-8B .002). On Gemma-3-12B the own transcript loses through return and budget states (+.121 [+.081, +.164] against the other's), and the other's through accepted wrong answers (+.115 [+.072, +.161] against the placeholder).

Table 20: **Own against another problem's redacted transcript.** Success with the placeholder (collected the same day), another problem's redacted failed transcript, and the own redacted transcript (original collection); the two contrasts with problem-paired bootstrap 95% intervals; the same-day placeholder minus the original placeholder; and the share of episodes whose accepted answer belongs to the donor's problem (post hoc). n: problems, 16 episodes per problem and condition.

Model	n	placeholder	transcript		own − other (primary)	other − placeholder (secondary)	placeholder new − original	answers the donor's problem
			other	own				
Llama-3.1-8B, 8k	130	.605	.557	.337	−.220 [−.264, −.175]	−.048 [−.074, −.023]	+.009 [−.015, +.032]	.005
Gemma-3-12B, 8k	85	.550	.465	.440	−.026 [−.090, +.033]	−.085 [−.127, −.042]	+.001 [−.024, +.026]	.071
Qwen3-8B, 16k	100	.682	.222	.607	+.385 [+.313, +.455]	−.460 [−.518, −.402]	+.001 [−.019, +.022]	.588
Qwen3-32B, 16k	99	.743	.514	.641	+.128 [+.078, +.176]	−.229 [−.272, −.188]	−.008 [−.024, +.011]	.242

Answer normalization. A scan specified after these results re-grades every frozen start of the mathematics start sets and the accepted wrong answers of the shared-controller mathematics conditions after a wider set of surface normalizations (option labels such as `\textbf{(C)}`, `\cfrac`, `{,}` thousands separators, tuple prefixes such as $(a, c) =$, spacing macros) and compares symbolic answers numerically. No answer graded correct differs numerically from the reference answer. Llama-3.1-8B and Gemma-3-12B have no frozen start that is correct under the wider normalization. The Qwen3-8B and Qwen3-32B shared-controller sets have 6 and 7 such problems of 108, where the model had boxed a correct answer with its option label or in a form the grader reads as a string, and the high-effort gpt-oss-120b set has 1 of 61; the other twenty start sets have none. On those problems the rejected class is the correct answer in one surface form. Registered analyses are reported unchanged; with those problems removed, $D_{\text{ret}} - D_{\varnothing}$ is −.126 [−.183, −.074] on Qwen3-8B (from −.120) and −.132 [−.181, −.089] on Qwen3-32B (from −.106), the primary contrast of Table 19 is −.078 [−.129, −.031] and −.120 [−.168, −.073], its other two contrasts still have intervals containing zero, and $C_{\text{full}} - A$ for gpt-oss-120b at high effort is −.103 [−.175, −.032] (from −.101). Relabeling alone, counting as successes the accepted answers that are correct under the wider normalization with no problem removed, gives −.067 [−.121, −.012] and −.100 [−.151, −.054] for that primary contrast, −.263 on Llama-3.1-8B, and no change on Gemma-3-12B.

A second domain. The same two policies were run on code: Llama-3.1-8B on 150 LiveCodeBench problems (Jain et al., 2024) at 8k. Starts were frozen from an existing pool of 64 draws per problem at 4k, two wrong, complete, fenced programs per problem; the rejected class is the start's test-outcome signature, grading is sandboxed execution, and a complete output without a fenced program is unparseable, a terminal failure. The truncation pilot passed (.023). Cell means are .142 for $D_{\varnothing}$ and .125 for D_{ret}, and $D_{\text{ret}} - D_{\varnothing}$ is $-.017$ $[-.043, +.009]$: not detected. In a split by the pool's success rate, registered before any data, both policies are at the floor on the 84 problems where no pool draw was correct (.000 and .002), and on the other 66 the contrast is $-.041$ $[-.099, +.019]$. The parse rule was not policy-neutral: .201 of $D_{\varnothing}$'s episodes and none of D_{ret}'s ended unparseable, and 60 of those 483 episodes had been graded correct by the sandbox; a plausible reason, not tested, is that the model returns bare code after a placeholder and fences its program when a fenced one is in context. Counting the 60 as successes, a lower bound on $D_{\varnothing}$ under a lenient rule, gives .167 for $D_{\varnothing}$ and a contrast of $-.042$ $[-.073, -.013]$ (post hoc). With the transcript in context the first call returns to the rejected class at .34 against .20 for $D_{\varnothing}$, and is correct at .085 against .119 (.142 under the lenient rule; Table 21).

Table 21: **The split of Table 2 for two further collections:** Gemma-3-12B with both policies rerun at a 16k ceiling on the same problems, and the code study (Llama-3.1-8B, LiveCodeBench, 8k). Columns, intervals and status as in Table 2.

	success			failure				
						other terminal failures		
Collection	first call	later calls	total $= J$	accepted wrong	budget	full ceiling	unparseable	call cap
Gemma-3-12B, 16k	$-.321$ $[-.388, -.259]$	$+.210$ $[+.162, +.257]$	$-.111$ $[-.168, -.054]$	$-.001$ $[-.063, +.063]$	$+.080$ $[+.043, +.121]$	$-.001$	$+.001$	$+.032$
Llama-3.1-8B, code, 8k	$-.034$ $[-.062, -.008]$	$+.018$ $[+.002, +.033]$	$-.017$ $[-.043, +.009]$	$+.188$ $[+.144, +.235]$	$+.031$ $[+.012, +.053]$	$+.001$	$-.201$	$-.002$

L THE POST-TRAINING STUDY

Design. The study asks whether training on retry prompts can reverse the ranking of the kept transcript and the same retry with the transcript withheld, and what it does to the level of success (§6). Llama-3.1-8B was fine-tuned on retry prompts with its own wrong solutions under three recipes and evaluated on a sealed test set of 256 problems $\times$ 8 episodes per policy (24,576 episodes over the four models). Training, validation and test problems come from MATH train, contamination-screened and near-duplicate-grouped, $1{,}024/128/256$; a problem enters only where the frozen base emits a complete, parseable wrong solution within eight draws, which excludes about 37% of candidates. A is a fresh draw; $D_{\varnothing}$, the withheld condition, and D_{ret} share one controller and differ in the retained turn and its instruction clause (§2, Appendix E), so $J = D_{\text{ret}} - D_{\varnothing}$ nets the exclusion out. The main set of five contrasts carries Bonferroni-corrected 99% intervals.

Recipes. All three train from the frozen base and export at 25%, 50% and 100% of training; LS and FS train on the same rows, in the same order, with the same loss mask, for two epochs at an effective batch of 16. LS is LoRA supervised fine-tuning (rank 32, scaling factor 64, dropout .05, on the attention and MLP projections, learning rate 10^{-4}, completion-only loss on the problem's reference solution). FS is full-parameter supervised fine-tuning on the same data and schedule at learning rate 10^{-5}, trained with 32-bit floating-point (fp32) master weights and exported as bfloat16 (bf16). LR is LoRA with online GRPO on 1,024 retry states, four candidates per state, sixteen groups per update, 64 updates, learning rate 5×10^{-6}, KL coefficient .01 against the frozen base, and a $0/1$ reward with no other term. FS and LR are one seed each; LS was trained with three seeds, and Table 5 reports the first. A checkpoint is selected per model by validation D_{ret} before the test set is unsealed (hardware in Appendix E).

Results. Table 5 gives the sealed test. The two supervised models give a positive J with A and $D_{\varnothing}$ below the base, and FS's D_{ret} is not separated from the base's. The remaining two members of the main set, with 99% intervals, are $J_{\text{LR}} - J_{\text{LS}}$ $-.255$ $[-.315, -.196]$ and $J_{\text{FS}} - J_{\text{LS}}$ $+.028$ $[-.016, +.073]$. Averaged over the three policies, LS moves $-.041$ $[-.068, -.014]$ and FS $-.208$ $[-.239, -.178]$ against the base, and LR $+.016$ $[+.004, +.029]$ while leaving J at $-.211$ and

ΔJ not detected ($+.014$ $[-.017, +.045]$). The two supervised models also shorten the fresh draw: a base episode of A emits 1,011 output tokens and truncates in .066 of episodes, against 319 and .007 for LS and 345 and .017 for FS, while LR is at 1,032 and .074.

Checks. A blind re-check of 160 graded items by a language-model assistant with the model and policy hidden agreed on $80/80$ items graded correct and relabeled $8/80$ graded wrong; no main-family point estimate moves by more than .0015 and no result row changes.

Further LS seeds. The LS recipe was rerun with two further training seeds (adapter initialization and sample order), everything else fixed; each seed's checkpoint was selected on validation D_{ret} by the same code before its test episodes were collected (for the third seed a rounding in that code departs from the rule; Appendix G), and the first seed was re-read with the same code (Table 22). J is positive with its 95% interval above zero in all three, and in all three the best tested policy stays below the base model's fresh draw.

Table 22: **LS across training seeds, sealed test.** 256 problems $\times$ 8 episodes, 95% problem-paired intervals. The last column is the best tested policy (D_{ret} in every row) minus the base model's fresh draw (.522).

Seed	checkpoint	A	$D_{\varnothing}$	D_{ret}	$J = D_{\text{ret}} - D_{\varnothing}$	$D_{\text{ret}} - A_{\text{base}}$
first	100%	.354	.374	.418	$+.044$ $[+.018, +.070]$	$-.104$ $[-.140, -.068]$
second	100%	.360	.373	.430	$+.057$ $[+.031, +.082]$	$-.093$ $[-.127, -.057]$
third	50%	.328	.330	.366	$+.037$ $[+.012, +.061]$	$-.156$ $[-.190, -.123]$

Mixed data. LS trains only on retry prompts. A control keeps the LS recipe and rows and adds, for every row, one direct-solution row: the same problem as a bare prompt with the same reference solution as target (2,048 rows, 256 updates, one seed; checkpoint selected on validation D_{ret} by the same rule, at half of training). The analysis and both publishable directions were registered before training. On the sealed test A, $D_{\varnothing}$ and D_{ret} are .379, .370 and .406 and $J = +.036$ $[+.015, +.059]$; against the base A is $-.143$ $[-.178, -.107]$, $D_{\varnothing}$ $-.116$ $[-.153, -.080]$ and D_{ret} $+.146$ $[+.105, +.185]$, beside LS's $-.169$, $-.112$ and $+.157$. The fall of A and $D_{\varnothing}$ is still there with direct-solution data mixed in one to one, and the best tested policy stays below the base model's fresh draw, $-.116$ $[-.150, -.082]$ (read outside the registration). End to end the mixed-data model is not separated from LS (Table 24; differences $+.009$ $[-.009, +.028]$, $-.001$ $[-.020, +.017]$ and $+.011$ $[-.008, +.030]$ for A, $D_{\varnothing}$ and D_{ret}).

Serving precision. To test whether FS's fall in A and $D_{\varnothing}$ is a rounding effect of bf16 serving, the base and FS were each served at bf16 and at fp32 on one H100 node and A and $D_{\varnothing}$ re-run on the sealed test set (256×8, fresh seeds). FS minus base is $-.350$ $[-.389, -.310]$ and $-.276$ $[-.313, -.236]$ at bf16 and $-.343$ $[-.379, -.307]$ and $-.281$ $[-.317, -.243]$ at fp32, for A and $D_{\varnothing}$. Within FS, fp32 minus bf16 is $+.002$ $[-.018, +.021]$ and $-.005$ $[-.025, +.016]$; within the base, $-.005$ $[-.029, +.017]$ and .000 $[-.023, +.022]$. No precision effect is detected, and no interval lies inside the $\pm.02$ band registered for calling the two precisions equivalent. The fp32 servers cannot carry the LoRA serving flags; at bf16 the base with and without those flags differs by $-.002$ $[-.025, +.021]$ and $+.009$ $[-.014, +.032]$.

Each model on its own errors. Table 23 repeats $D_{\varnothing}$ and D_{ret} on wrong starts each model generates for itself. The supervised models produce a parseable wrong start on more of the 256 problems than the base does; within a row, the base and the GRPO model keep a negative J with the interval excluding zero, and for the two supervised models J is not detected.

End to end. The sealed test and Table 23 condition on a failure. The end-to-end quantity does not: for retry policy φ, $\mathcal{E}_{\varphi}$ is the probability that the model's first attempt is correct, plus the probability that it is complete, parseable and wrong times the success of one φ episode started from that attempt. The problem list is 400 test-split problems that no failure filter has touched: the rows of the candidate list after the last one scanned when the failure set was built, disjoint from the training, validation and test sets. Each model makes 8 first attempts per problem at a 4,096-token ceiling; every complete wrong attempt starts one episode per policy under the evaluation controller at 8,192, so the three policies share a total ceiling of 12,288, and A here is a fresh draw. A first attempt that is truncated or unparseable scores zero under every policy. Quantity, problem list and contrasts were

Table 23: **Each model on its own wrong starts.** Coverage is the fraction of the 256 test problems on which that model emits a complete, parseable wrong start within eight draws. 95% problem-paired intervals.

Model	coverage	$D_{\varnothing}$	D_{ret}	$J = D_{\text{ret}} - D_{\varnothing}$
base	.820	.419	.237	$-.182\ [-.229, -.136]$
LS	.945	.348	.374	$+.026\ [-.012, +.064]$
LR	.824	.399	.239	$-.160\ [-.209, -.111]$
FS	.973	.177	.169	$-.008\ [-.042, +.026]$

registered before collection; intervals are problem-paired bootstrap, 95% (Table 24). Carrying the transcript, $\mathcal{E}_{D_{\text{ret}}} - \mathcal{E}_{D_{\varnothing}}$, is below zero with the interval excluding it at base and on LR and is not detected for LS and FS (last column of Table 24). The withheld condition against the fresh draw, $\mathcal{E}_{D_{\varnothing}} - \mathcal{E}_{A}$, is not detected at base ($-.004\ [-.014, +.005]$), for LS ($+.012\ [-.001, +.024]$) or for FS ($+.008\ [-.005, +.021]$), and is $-.016\ [-.026, -.007]$ on LR. Across models the ordering of the A and $D_{\varnothing}$ columns of Table 5 holds under every policy, D_{ret} included: against the base under the same policy, LS is $-.135\ [-.161, -.110]$, $-.119\ [-.145, -.094]$ and $-.086\ [-.112, -.060]$ for A, $D_{\varnothing}$ and D_{ret}, FS is $-.366$, $-.354$ and $-.327$ with intervals excluding zero, and LR's three intervals contain zero. The first attempt carries most of it: LS is correct on .460 of first attempts against the base's .650, FS on .268.

Table 24: **End to end: first attempt, then one verified retry episode** (400 unfiltered problems $\times$ 8 first attempts; total ceiling 12,288). Retry success is over the model's own complete wrong first attempts; the remaining first attempts, truncated or unparseable and scored zero, are .087, .085, .046, .120 and .039 in row order. LS: first seed.

	first attempt		retry success			end to end			
Model	correct	wrong, startable	A	$D_{\varnothing}$	D_{ret}	$\mathcal{E}_{A}$	$\mathcal{E}_{D_{\varnothing}}$	$\mathcal{E}_{D_{\text{ret}}}$; $\mathcal{E}_{D_{\text{ret}}} - \mathcal{E}_{D_{\varnothing}}$	
base	.650	.263	.340	.323	.194	.739	.735	.701;	$-.034\ [-.045, -.024]$
LR	.655	.260	.355	.292	.178	.747	.731	.701;	$-.030\ [-.039, -.020]$
LS	.460	.494	.292	.316	.314	.604	.616	.615;	$-.001\ [-.014, +.012]$
FS	.268	.612	.172	.185	.174	.373	.381	.374;	$-.007\ [-.021, +.006]$
LS, mixed data	.473	.488	.289	.291	.314	.614	.615	.626;	$+.011\ [-.001, +.024]$