IT HAS THE ANSWER AND KEEPS GOING: HOW FINE-TUNING BREAKS THE DECISION TO STOP

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning on long reasoning traces is how most open reasoning models are built, and its students often fail to finish: ours runs to the budget on half of its AIME 2025 generations. The cause is a learned stop decision, which fine-tuning can break while leaving reasoning intact. Serve a Qwen3-8B student without its eleven-token, task-free system turn and its computed accuracy moves 0.6 points while its delivered accuracy falls 29.4: it boxes the correct answer early and keeps going. One principle explains every failure we find: *a distilled student stops only under conditions that accompanied every stop it was shown*. Three causes follow, each confirmed by controlled intervention. **(i)** An end-of-turn token dropped from supervision is silenced ($81.7\% \rightarrow 0.0\%$ of generations) while a standard harness scores both checkpoints alike. **(ii)** A stop tied to the training prompt fails at the two closing tokens; editing their probabilities restores termination, and training on two formats prevents it. **(iii)** Every demonstration stops on a converged answer, so a failing student, whose candidate answers never settle, stops after only 23.5% of its wrong answers (base: 96%), a pattern 17 of 21 released long-CoT models share. Trained on stops alone, the student collapses; mixed with original data, a moderate dose halves wrong-and-capped generations at no significant accuracy cost on one training seed, and stronger doses abandon solvable problems. TERMINATOR, one guard per cause, terminates non-termination at inference: on every student we trained and benchmark we tested, capping falls to zero and delivered accuracy rises (4.6 points under the training format, 21.7 without the system turn; also pooled over 20 released checkpoints) while computed accuracy never falls; budget forcing, which also forces a commitment, does as well, as the principle predicts. Reported termination is partly a property of the harness: four standard defaults span 13.4 points for one checkpoint.

1 INTRODUCTION

Distilling long reasoning traces from a strong teacher is the standard recipe for open reasoning models (DeepSeek-AI, 2025; Guha et al., 2026; Muennighoff et al., 2025; Ye et al., 2025b), and its students often fail to finish. Our Qwen3-8B student runs to its token budget on half of its AIME 2025 generations and a quarter of its GPQA-Diamond ones; 17 of 21 testable released long-CoT checkpoints share its signature; and a popular trainer's defaults would cut the terminal token from 99.8% of our training episodes. The literature calls budget exhaustion "overthinking", a property of the model, and answers it at decoding time with length penalties, budget forcing or early exit (Chen et al., 2024; Sui et al., 2025; Aggarwal & Welleck, 2025; Yang et al., 2025b).

We find instead that fine-tuning separates two capabilities that evaluation conflates: computing an answer and delivering it. Delete the eleven-token system turn every training example carried and the student's computed accuracy (any boxed answer) moves **0.6 points** while its delivered accuracy (answers given after it stops) falls **29.4**; on the generations it now fails to finish, it had boxed an answer, 97% of them correct, a median 11% into the budget. A harness that extracts answers from anywhere credits it; the user never receives it.

One principle explains where the decision breaks: a distilled student stops only under conditions that accompanied every stop it was shown. For closing the thought and ending the turn, these conditions are a supervised end-of-turn token, the training system turn and a converged

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

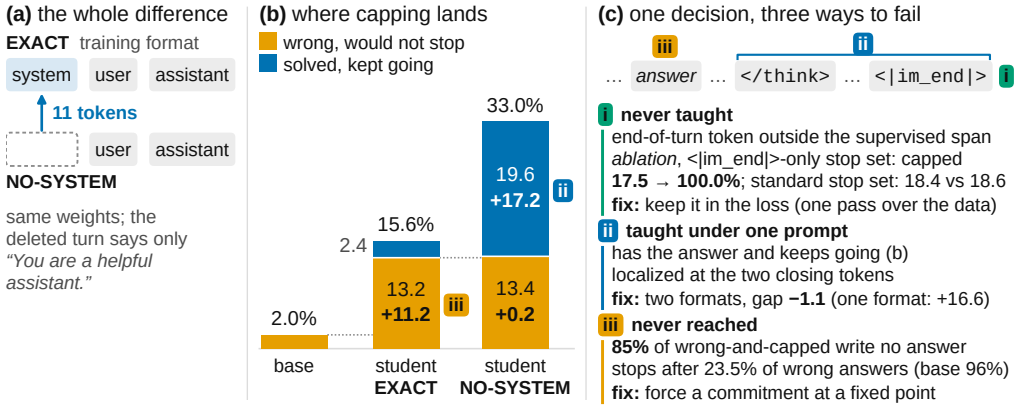

Figure 1: **The student computes the answer but does not deliver it: one decision—closing the thought and ending the turn—fails three ways: (i) always, (ii) after solving, (iii) after a wrong answer.** **(a)** The two prompts differ only by the eleven-token system turn. **(b)** Capping (hit the budget) = wrong × (1 − stop-after-wrong) (orange) + solved-then-capped (blue); bold: change vs. left bar. Main student, MATH-500, one decoding seed (two: gap +16.6). **(c)** Each cause at the end of a generation, and its fix (fixed point: token position).

answer; each missing condition is one cause (Table 1). Capping (hitting the budget) splits as cap = wrong × (1 − stop-after-wrong) + solved-then-capped into a wrong branch (wrong, and never stopped after) and a solved branch (solved, then kept going); (i) fills both, (ii) the solved one, (iii) the wrong one (Figure 1). Separating the causes is hard: all three end at the same two tokens, produce the same capped generation, and pass a harness that scores computed accuracy. We isolate each with one-variable training ablations across eight initializations, a causal edit of two logits during decoding, pre-registered dose sweeps against a matched sham, and four benchmarks.

(i) Never taught. Truncation, packing or a loss mask can drop the terminal token from supervision (from all 287,735 assistant turns of one agent-trace set of ours). Cutting it silences the token, yet a standard harness scores both checkpoints alike because the base model's end-of-text token takes its place (§3); counting episodes whose last supervised token is terminal catches it in seconds.

(ii) Taught under one prompt. Exposure bias (Ranzato et al., 2016; Ross et al., 2011; Arora et al., 2022) and template overfitting (Lin et al., 2025) predict damage spread over the reasoning. Instead, in the 8B and 7B students the two formats' predictions diverge 100–1000× more at the two closing tokens than across the thought, and editing those two tokens' probabilities mid-decode restores stopping (Figure 2); in five brittle students the solved branch carries 64–97% of the gap. Two formats from step 0 prevent it at no accuracy cost (Figure 3), and 86 rows in the other format repair it after the fact (§§4, 6).

(iii) Never reached. The 30,000 demonstrations close a median 45 tokens after the teacher's correct answer; the dataset's judge, our label check and our length cut removed every search that did not converge. A failing student holds candidate answers as accurate as its own resamples, but they never settle, and it stops after 23.5% of wrong answers against its base's 96%. Stop supervision fixes it only in part: a 100× sweep of terminal loss moves nothing, and stops trained alone from answer-less states collapse the student. Mixed with replayed original data, a moderate dose halves the wrong branch with no significant accuracy cost, and stronger ones abandon solvable problems, because its hidden state tells a failure from a slow success barely better than elapsed position does. Training moves a threshold; a decode-time commitment removes the rest (§5).

TERMINATOR terminates non-termination at inference and prevents two of its causes in training (Table 1). A data check and two formats of every example remove causes (i) and (ii) before training; a decode guard that, at a fixed position, delivers a boxed answer or asks for one, brings capping to zero on every student and benchmark we tested, raising delivered accuracy by 4.6 points under the training format and 21.7 without the system turn, in every AIME 2025 and GPQA-Diamond cell and on 20 released checkpoints pooled, with no significant accuracy loss (Table 4); budget forcing (Muennighoff et al., 2025) does as well, as the principle predicts: the stuck student lacks a commitment, which more thought does not supply.

Table 1: **TERMINATOR: one test and one guard per cause.** Each test needs only artifacts at hand (training data, checkpoint, base), no annotation; §§3–5 establish the rows. Stop-after-wrong: rate of stopping after a wrong answer.

cause	test, on artifacts you have	guard, what removes it	effect (cost)
(i) end-of-turn token never supervised	count episodes whose last supervised token is not the end-of-turn token	DATA GUARD: keep the end-of-turn token in every supervised span	**0%** of generations end on it without, **81.7%** with (seconds, no model)
(ii) stops only under the trained prompt	score two formats that differ only in the system turn	FORMAT GUARD: a second format from step 0; after the fact, 86 reformatted rows	gap **+16.6** $\rightarrow$ **−1.1** (no extra steps)
(iii) stops only after a converged answer	compare own-format stop-after-wrong with the base's	DECODE GUARD: at a fixed position, deliver the held answer or force one; training halves it	capping $\rightarrow$ **0**, delivered **+4.6** (no training)

A reported termination number is thus partly a property of the harness: with only the prompt bytes differing, one checkpoint spans 13.4 points across four standard defaults and TinyR1-32B-Preview 32.2; evaluations should report delivered accuracy under two prompt formats.

2 SETUP: ONE RECIPE, EIGHT INITIALIZATIONS

Eight students. We fine-tune Qwen3-8B-Base (Yang et al., 2025a) (main student; plain cross-entropy), Qwen2.5-7B, DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI, 2025), Llama-3.1-8B and -Instruct (Grattafiori et al., 2024), Mistral-7B-v0.3 and, as pre-registered blind tests, Qwen2.5-14B and DeepSeek-R1-Distill-Llama-8B for 3 epochs on the same 30,000 long-CoT mathematics traces from Nemotron-Math-v2 (Du et al., 2025) (gpt-oss-120b; median thought 7.8k tokens, p90 16.6k; all closed, none truncated); §4 adds science and code students. Every example is `system: "You are a helpful assistant."` + bare problem + target beginning `<think>`.

Six byte-exact formats. A *format* is the byte-exact prompt serialization, the token string the chat template emits for a problem. EXACT is the training format byte for byte; NO-SYSTEM drops the system turn; INSTR adds the NeMo-Skills instruction line; HISTORICAL is NO-SYSTEM+INSTR, that harness's default; two more, ALT-SYS and NO-SYS+ALT, appear in no training set (Appendix A12.1).

Two accuracies, split by branch. vLLM (Kwon et al., 2023), T=0.6, top-p=0.95, fixed seeds, prompts as exact token ids; budget 8,192 tokens for MATH-500 and 30,000 for AIME 2025 (n=360 for the main student). A generation is *capped* when it reaches the budget, otherwise a *well-formed* (thought closed) or *unclosed* stop (tables: *malformed*); a capped generation is solved-then-capped if its last boxed answer is correct and *wrong-and-capped* otherwise; *premature* (no boxed answer) and *loop* are flags. Comparisons are paired by (problem, seed); a difference is *significant* when its 95% CI (problem bootstrap) excludes zero. *Computed* accuracy reads the last `\boxed{}` anywhere, *delivered* only after a closed thought and a natural stop; *stop-after-wrong* is $P(\text{natural stop} \mid \text{last box wrong})$, the rate of stopping after a wrong answer. The identity's two terms are the *solved* and the *wrong* branch. Cause (iii) is present when a student caps more than its base under its own format (its training or documented format), beyond decode noise, with stop-after-wrong below the base's. A student is *brittle* (a pre-registered endpoint) if deleting the system turn, or adding a generic one where the template has none, raises capping significantly and by more than 5 points. The main student's format gap (NO-SYSTEM−EXACT capping) is +16.6 (two decoding seeds); capping varies up to 3.2 points across decodes, so each number states its condition.

3 CAUSE (I): AN UNSUPERVISED END-OF-TURN TOKEN IS NEVER EMITTED

The terminal token ends the supervised span and falls outside it whenever truncation or packing cuts a long example or a loss mask skips it. The three packing stacks we read guard the cut by hand, no two alike, none verifying its output; Li et al. (2026) trace a 46.8% BBH regression to

Table 2: **Cutting the end-of-turn token silences it (81.7% to 0%); a standard harness's stop set sees nothing.** Only the terminal token's supervision differs. MATH-500, EXACT, 500×4 seeds (left), 500×2 (right).

% of generations	stop set of a standard harness				stop only on the end-of-turn token			
end-of-turn	capped	well-formed	acc.	on <|im_end|>	capped	well-formed	acc.	on <|im_end|>
supervised	18.4	68.7	79.2	81.7	17.5	70.8	78.3	82.5
unsupervised	18.6	81.4	79.6	0.0	100.0	0.0	45.5	0.0

one masked end-of-sequence loss in continued pre-training, and practitioners report the symptom unmeasured (axolotl-ai-cloud contributors, 2025; hxs91, 2024). In one of our trace sets (8,858 episodes, 6,496 agentic-coding sessions) a loss mask left *every* assistant turn's terminator unsupervised while 97.3% end in a completed tool call: a model trained on it learns to close every structure it opens but never to stop. Our mathematics students keep it only because ms-swift drops over-length rows; TRL's defaults would cut it from 99.8% of episodes. Under softmax cross-entropy each supervised non-terminal position lowers the end-of-turn token's logit in proportion to its probability, and only the terminal target (one per 9,839 body tokens) raises it; without that target the logit gets only downward gradient, so silence is optimal, as Table 2 confirms.

The cut arm emits 0 of 2,000 end-of-turn tokens, unseen by the harness. Under a standard harness's stop set both arms read alike on capping, accuracy and loops (none significant), yet the cut arm never emits the end-of-turn token (0 of 2,000): it falls back on the base's <|endoftext|>, which that stop set accepts. Stopping only on that token, it caps every generation, solved or wrong: cause (i) fills both branches. Cause (ii) also silences it: without the system turn the supervised arm stops emitting it ($81.7 \rightarrow 1.1\%$). The check is one data pass: refuse to train unless each episode's last supervised token is terminal, which separates the arms completely in seconds, with no model.

4 CAUSE (II): DELETING THE SYSTEM TURN BREAKS THE STOP

Deleting the training system turn keeps solving and breaks stopping: in a 2×2 factorial of system turn and instruction line, accuracy stays at 79.8–84.6% while solved-then-capped generations (solved but run to the budget) rise eightfold and loops seventeenfold; a second decoding seed, a 30,000-token budget and another student reproduce it. The turn's presence carries it: an empty five-token turn recovers 89% of the capping gap, an unrelated sentence 64% (tables: Appendix A9).

Computed accuracy moves 0.6 points; delivered accuracy falls 29.4. Under NO-SYSTEM, 103 of the 165 capped MATH-500 generations box a correct answer and run a median 7,107 tokens past it without closing; under EXACT only 12 of 78 do. Over both seeds such generations loop verbatim ("Now produce final answer.") and close the thought in 2.7%; 34.6% of all tokens come *after* the answer, against 2.8%. Delivered accuracy (one seed) falls from 80.2 to 50.8%, -29.4 points $[-34.0, -25.0]$, against 0.6 for computed accuracy, which our judge and standard harnesses report; the natural stop alone costs 17.8. TinyR1-32B-Preview, a released single-dataset distill, is 92.0 computed and 61.6 delivered under the public simple-evals default, unlike every multi-format checkpoint we scored; Lotfi et al. (2026) see the same in quantized reasoners.

The damage sits at two closing tokens; restoring them is enough. Held to the teacher's own trajectories, the student's probability of ending the turn at the terminal state falls by a quarter (median $0.975 \rightarrow 0.739$), while mid-thought it is *equally* reluctant under both formats. Teacher-forced on held-out traces the divergence is two to three orders of magnitude larger at the two closing tokens than over the reasoning (Figure 2a). To test causality we edit only those two tokens' probabilities: with p_t^{EXACT}, $p_t^{\text{NO-SYSTEM}}$ the temperature-scaled next-token distributions given the *same* generated history, S the two closing tokens and $p(S) = \sum_{u \in S} p(u)$, we sample from

$$q_t(v) = p_t^{\text{EXACT}}(v) \;\; (v \in S), \qquad q_t(v) = \frac{1-p_t^{\text{EXACT}}(S)}{1-p_t^{\text{NO-SYSTEM}}(S)}\, p_t^{\text{NO-SYSTEM}}(v) \;\; (v \notin S), \tag{1}$$

so the two closing tokens take their training-format probabilities and all other ratios are unchanged. Figure 2b: restoring them recovers training-format behavior; the same edit on other tokens that move

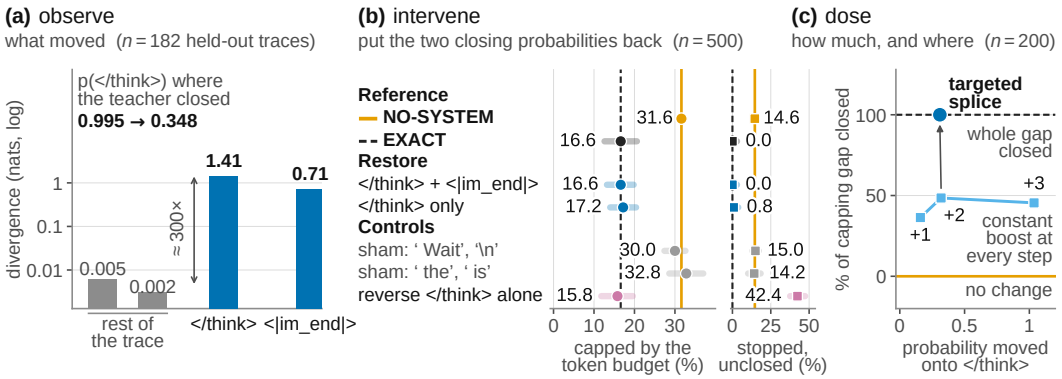

Figure 2: **Two closing tokens carry the failure: the formats diverge most there, and restoring only them undoes it.** **(a)** Next-token divergence, EXACT (training prompt) vs. NO-SYSTEM, on held-out teacher traces (nats). **(b)** MATH-500, NO-SYSTEM with named tokens at EXACT probabilities (Eq. 1); circles capped, squares unclosed; 95% CIs. **(c)** Gap closed vs. probability moved.

more has no effect; reversing them reproduces the failure and overshoots it, so the other tokens' probabilities also contribute. Figure 2c: termination returns when closing probability is added *at the right step*. Applied only at steps where the training format would itself close, the edit rescues fully while changing 0.8 steps per generation; a constant logit bias on the same tokens saturates near half the gap while moving three times the probability. The rescue replicates on a second training seed, the science and the Qwen2.5-7B students, so decode-side remedies work downstream of the cause.

A base-side screen flags brittle initializations before training. Each closing token is supervised once per episode against a mean 9,839 body tokens, so it breaks when the deleted prefix supplies much of its logit margin, a share readable from the *base* before fine-tuning: D, the shift the deletion applies to the base's closing logit on held-out traces, separates the six initializations we trained ($D \geq 1.97$ where students broke, ≤ 0.85 where they held) and reproduces on a disjoint cohort. Two blind tests separate D from model family: Qwen2.5-14B, predicted safe, held ($+0.2$, two decoding seeds); R1-Distill-Llama-8B, predicted brittle, moved the predicted way but short of the endpoint.

A second format erases the 16.6-point gap. Two pre-registered arms that differ from the main student only in formatting make the link causal (Figure 3). Formatting half of the rows with the system turn and half without, from step 0, removes the dependence (-1.1 $[-3.0, +0.8]$) at unchanged EXACT capping and accuracy, no held-out format worse; so does the same training on Qwen2.5-7B, on the science student (the most brittle we trained), with the 15% wrong-labeled traces of §5 added, and on two more training seeds. On GPQA-Diamond and AIME 2025 no two-format student shows a gap (-2.5 to $+1.9$, against $+7.1$ to $+45.3$ single-format). The deletion also lowers stop-after-wrong in four of the five brittle students (Qwen2.5-7B 27.5 → 12.4%) and in no two-format student: two formats protect both branches. The mirror arm falsified our prediction: a student trained with *no* system turn does not break when a generic one is added (Figure 3, bottom), because the added turn removes 0.38 of its closing margin where deletion removes 0.90 of the main student's (as for safety prompts, Lyu et al., 2024b). The dependence is on *deleting* what was seen.

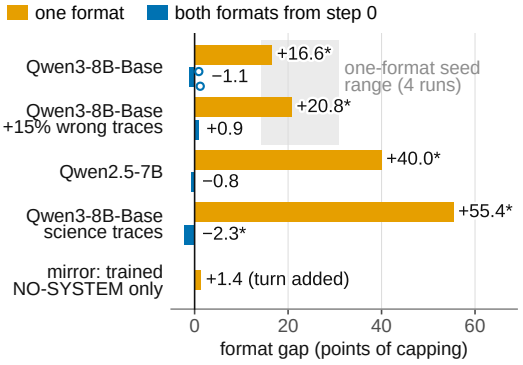

Figure 3: **Trained on both formats from step 0, none of four students breaks; trained on one, all four do.** NO-SYSTEM−EXACT capping, MATH-500, two decoding seeds (science one-format: one). Gray: four one-format training seeds; circles: two more two-format training seeds (123 final, 7 after epoch one). * significant.

fine-tuned by us	MATHEMATICS		SCIENCE	CODE	
maths traces unless noted		AIME	GPQA	LCB	LCB
EXACT → NO-SYSTEM	MATH-500	2025	Diamond	stdin	starter
Qwen3-8B-Base	**+17.4**	**+17.5**	**+28.0**	**+31.7**	**+18.9**
2nd training seed	**+29.0**	**+17.5**	**+33.2**	**+24.6**	**+23.9**
science traces	**+55.4**	**+58.3**	**+45.3**	**+65.0**	**+75.0**
code traces	**+4.8**	**+7.5**	**−6.1**	+0.2	**+8.2**
Qwen2.5-7B	**+36.4**	+7.1	**+13.9**	**+22.8**	**+12.5**
R1-Distill-Qwen-32B	**+39.6**	**+30.8**	**+10.5**	**−7.8**	+3.2
Llama-3.1-8B	+3.2	+2.5	**+4.2**	**+7.4**	**+11.1**
Llama-3.1-8B-Instruct	+0.2	−2.5	+1.0	**−7.8‡**	−7.5‡
Mistral-7B-v0.3	+2.6	+0.8	−0.9	+0.4	−1.4
Qwen2.5-14B †	+0.8	*n/a*	*n/a*	*n/a*	*n/a*
R1-Distill-Llama-8B †	+2.8	*n/a*	*n/a*	*n/a*	*n/a*

released students	MATHEMATICS	
	MATH-500	AIME 2025
NO-SYSTEM (own system turn removed)		
OpenThinker-7B	−1.0	+0.8
OpenMath-Nemotron-7B	**+2.9**	+0.8
R1-Distill-Qwen-32B	**+2.2**	−2.5
+generic system (template has none)		
OpenReasoning-Nemotron-32B	−0.2	+4.2
R1-Distill-Qwen-1.5B	−1.2	+4.2
R1-Distill-Qwen-7B	+0.6	**−5.8**
R1-Distill-Qwen-14B	**−1.4**	+4.2

+8.2 significant +0.8 not significant
n/a not evaluated ‡ reasoning off † **blind**

Figure 4: **Brittleness tracks family: every single-format Qwen-derived mathematics student breaks but the blind-predicted Qwen2.5-14B,** no Llama or Mistral one on mathematics or science (the Llama-3.1-8B base one on code), and no released student under deletion. Paired change in capping (ran to the budget; points) between the training format and one differing only in the system turn; one decoding seed, n=500 per MATH-500 cell, four seeds elsewhere. Bold: significant; breaking: $> +5$ and significant. †: blind students, first seed; R1-Distill-Llama-8B re-run, corrected tokenizer. ‡: reasoning switched off (54 of 728 code generations open a thought).

No base we probed carries the dependence; fine-tuning creates it. Qwen3-8B-Base and Qwen2.5-7B, whose students break, are insensitive to the deletion (Δcap -0.5 and $+1.0$, n=200 each); the only primary bases that move, Llama-3.1-8B and Mistral-7B-v0.3, move the opposite way, and their students held. On mathematics and science brittleness tracks model *family*, whether base or instruct (Figure 4); on code the line blurs, a family×corpus interaction that tokenization does not explain.

Why released models rarely show it. The deletion needs a student built from a raw base *and* served with a system turn; across 31 released checkpoints the two nearly exclude each other (Fisher exact p=0.023), the deletion test is well posed on four, and none of them breaks. The models whose looping is documented (Qwen3-8B/14B/32B and the R1 distills, Duan et al., 2026; QwQ-32B and s1.1-32B, Fan et al., 2025) are all in that panel, and no system-turn change moves them past $+2.2$ on MATH-500 under their documented formats: those reports measure cause (iii). TinyR1-32B-Preview, which breaks under an *added* generic turn, swings from 1.0 to 33.2% capped across public harness defaults; its parent spans 1.2.

One optimizer setting beats 16.5 points of seed noise. Training seed moves the format gap by 16.5 points across four runs, all brittle; one setting exceeds that noise. **Setting AdamW's two momenta equal (BalancedAdamW) removes most of the brittleness and leaves training-format behavior untouched**: across six matched pairs perturbed-format capping falls by 4.2 to 31.3 points (median 25) and own-format capping moves at most 1.9 (one-sided sign test p=0.016).

5 Cause (iii): every stop followed a converged answer

Cause (iii) is a low stop-after-wrong rate under the student's own format, which sets the wrong branch. Fine-tuning cuts stop-after-wrong fourfold and raises the wrong-and-capped term sevenfold; a harder problem or a short budget does not explain it: with 3.7× the budget, 60% of the stuck generations are still capped without a correct answer. Accuracy only scales this branch: swapping stop-after-wrong rates, the student would cap on 3.0% and the base on 34% (Appendix A6).

The student never reaches the stop: 85% of its failures write no answer. Of its 127 own-format wrong-and-capped generations, 85% write no `\boxed{}` answer, 92% keep re-deriving with varying text, and the thought closes in 0.8%. Among its wrong generations it stops after 66% of those that box an answer and 1.8% of those that do not; the base stops after 98 and 94% (Figure 5a). The decision is intact: fed teacher traces that end wrong and matched correct ones, it closes where the teacher closed as readily (median p(`</think>`) 0.995 in both; Qwen2.5-7B alike; pre-registered). The signature generalizes: in all eight base–student pairs whose stop-after-wrong drops, 76–92% of wrong-and-capped generations write no answer and the student stops almost only

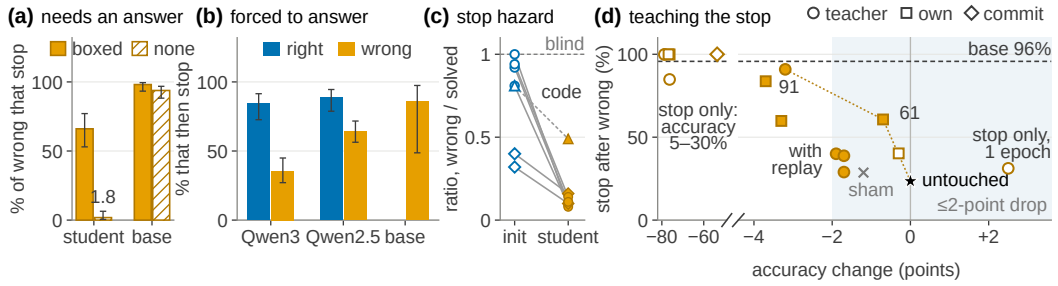

Figure 5: **A failing student stops only once it holds an answer (66% against 1.8%); stops taught without one collapse or trade stop-after-wrong for accuracy. (a)** Stop rate of wrong generations with and without a `\boxed{}` answer. **(b)** Stuck generations (no stop, no answer by 4,096 tokens) forced to close and answer: turn ended after right and after wrong answers. **(c)** Position-matched stop hazard, wrong over solved (1: no discrimination), in the eight pairs whose stop-after-wrong drops; diamonds: R1-distilled initializations; triangle: code student. **(d)** Fine-tuned on give-up rows from the teacher's (circles) or its own (squares) answer-less states, alone (hollow) or with replay (filled); diamond: distilled forced commit; dotted: frontier. MATH-500, EXACT (training prompt), budget 8,192; 95% CIs (Wilson).

Table 3: **Format fixes erase the 16.1-point solved-branch gap; none of these moves the own-format wrong branch (12.7; arms 12.1–13.2).** Qwen3-8B-Base student, MATH-500, budget 8,192; format gap and Δ branch: NO-SYSTEM−EXACT; n=1,000 (base: 500); * significant; § own baseline (untouched 15.7 / 30.9); † 97% correct; per arm: Appendix A5.

	own format (EXACT)					system turn removed (NO-SYSTEM)			
% of generations	capped	= wrong-and-capped	+ solved-then-capped	acc.	stop-after-wrong	format gap	Δ solved branch	Δ wrong branch	acc.
base model (no fine-tuning)	1.8	1.8	0.0	55.2	96.0	−0.4	–	–	62.6
main student (one format)	15.0	12.7	2.3	83.4	23.5	+16.6*	+16.1*	+0.5	82.8
aimed at cause (ii): the format									
both formats from step 0	16.4	13.0	3.4	82.6	25.3	−1.1	−0.7	−0.4	83.3
repair D, one epoch after the fact§	14.8	12.1	2.7	83.2	28.0	+1.6	−0.4	+2.0	81.4
terminal-token weight ×16 (control)	15.2	12.9	2.3	82.7	25.4	+0.8	–	–	82.7
aimed at cause (iii): stops shown after an answer, share of rows									
give-up after the answer, 12.9%	14.9	12.3	2.6	83.1	27.2	+48.8*	+43.3	+5.5	79.1
traces labeled wrong†, 15%	15.8	12.8	3.0	81.4	31.2	+20.8*	–	–	–
traces labeled wrong†, 30%	15.5	13.2	2.3	82.4	25.0	+23.6*	–	–	–
both formats + the 15% traces†	15.7	12.9	2.8	82.8	25.0	+0.9	−0.1	+1.0	–

once solved (Figure 5c), as do 17 of 21 testable released long-CoT checkpoints. Greedy decoding turns 81% of the branch into the verbatim loops of Pipis et al. (2026); at temperature 1.0 none remains, yet the branch is still 15.8% and stop-after-wrong 18.6% (base 98.0%): loops are one form of the persistence that studies of overthinking and long wrong answers measure (Duan et al., 2026; Fan et al., 2025; Fatemi et al., 2025; Chen et al., 2026).

A failing student holds 3.3 answers but never settles on one. Since the demonstrations close just after the teacher's correct answer (§1), the principle predicts cause (iii): stopping is conditioned on a converged answer, which a failure never holds. Across corpora the prediction holds: only the math corpus closes its thoughts on a stated answer (all but 3 of 30,000) and only its student never closes otherwise (0 of 852), while the science and answer-only corpora close without one and their students stop after 91–100% of wrong answers. A failure does hold candidates, but they never settle: forced to answer at ten points along a stuck generation, the student gives 3.3 distinct answers, each as accurate as its own resamples, and returns to an abandoned one in 52% of generations (12.5% when it solves). Appending its own commit line closes a solved generation (median p 0.99) but not a stuck one (0.001): stopping needs a settled answer. Forced to close and answer, the student ends the turn after 84% of right answers and 36% of wrong ones (Figure 5b; Fisher p=10^{-9}; the base after 83–100% of wrong ones). The principle also predicts which fixes fail: stops shown after an answer strengthen closing only where it already happens. Give-up sentences after the teacher's answer appear in 1 of 1,000 own-format generations and in no failure; traces labeled wrong-ending

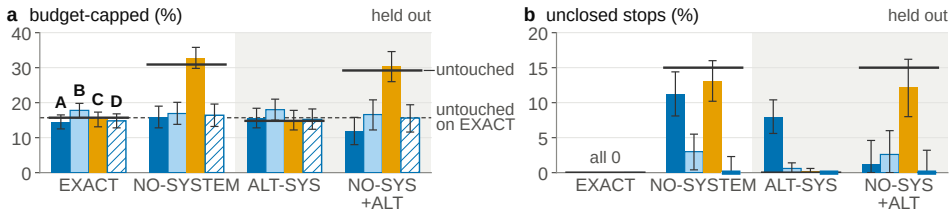

Figure 6: **A second format repairs the student in one epoch, without its own rollouts.** C (replay only) stays at the untouched level (ticks); A, B, D bring NO-SYSTEM (system turn removed) capping within 1.2 points of the training-format level (dashed) and transfer to held-out formats (shaded); A pays in unclosed stops (b). MATH-500, n=1,000 per bar (held out: 500); paired 95% CIs.

(97% correct) and a 100× sweep of terminal supervision move neither branch (Table 3; full sweep: Appendix Table 11).

Trained stops halve the wrong branch; pushed further, they cost accuracy. A failure holds no answer, and neither does a success still in progress, so a stop shown from an answer-less state has nothing to tell them apart by. Trained alone, the stop fires everywhere: five stop-only arms stop on almost every problem and keep 5–30% accuracy. Anchored by replay or kept to one epoch, it becomes a trade-off (Figure 5d). From 800 of the student's own stuck states it raises stop-after-wrong 11.6 points over a matched sham [+3.2, +20.4] at unchanged accuracy, a fifth of the wrong branch. A moderate dose passes our pre-registered MATH-500 test: it halves the wrong branch (12.7 → 6.8), cuts capping from 15.0 to 8.3% there, 49.7 to 11.4% on AIME 2025 and 27.0 to 6.6% on GPQA-Diamond, with no significant accuracy change (−0.7 [−2.5, +1.1]; GPQA −2.8 [−5.8, +0.3]; Table 4). With 2,000 teacher cuts over three epochs stop-after-wrong returns to 90.9% (base 96%) and the wrong branch to 1.8, but the give-up sentence is only twice as likely in a failure as on a path to a solution (0.06 against 0.03 per paragraph), so 40% of give-up stops abandon solvable problems and accuracy drops 3.2 points (Qwen2.5-7B: 18.2 → 0.8, −2.0). A gate trained on the value of continuing from the same state separates the two barely better than elapsed position when it reads the student's hidden state (held-out AUC 0.65 against 0.64), and well when it reads whether the student's forced answer has settled (0.83, exploratory), the property every demonstration had. A decode guard at a fixed position reads the budget instead: an answer prompt there makes the stuck student commit in 94–98% of cases, at the accuracy of its own resamples (22%). Causes (i) and (ii) are gaps in the data and close when the data are fixed. Cause (iii) needs a stop conditioned on an unsettled failure, which no demonstration of a converged answer contains; training still recovers part of it: averaging the moderate-dose weights with the untouched student raises delivered accuracy on MATH-500 by 2.5 points [+0.9, +4.1], and on GPQA-Diamond and LiveCodeBench, with no decode-time help (exploratory, one training seed; Appendix A6); the decode guard removes the rest.

6 TERMINATOR: ONE GUARD PER CAUSE, NO ACCURACY COST

The data guard is the check of §3, the format guard two formats from step 0 (Figure 3) or one epoch of reformatted replay; the decode guard, at a fixed token position T, closes a still-running generation and delivers its boxed answer or prompts for one.

The decode guard lifts delivered accuracy everywhere; so does budget forcing. At T (pre-registered on one training-format seed) it fires on 5.8% of own-format generations holding an answer, 64% of them correct, and on 13.0% holding none, whose forced answers are right 31% of the time against 11.5% if left to run. It raises delivered accuracy on every student and benchmark and costs accuracy in none (Table 4), restoring all of the deleted turn's cost on the 32B student (+36.7 → +1.0) and, on 20 released checkpoints under their own formats, significantly in 8. Any forced commitment works as well: an answer-agnostic prompt at T delivers 0.6 points more, and s1's budget forcing (Muennighoff et al., 2025) at the full budget 1.6 more for 340 more tokens. The data guard stays necessary: the end-of-turn-cut student, which never emits the end-of-turn token, still trails its twin (Table 32).

Table 4: **Every guard removes its failure at no accuracy cost (decode: +2.2 to +44.7 delivered); give-up training cuts capping in every cell, at no significant accuracy cost only from the student's own states.** Change vs. untouched (decode: delivered acc.); gap: NO-SYSTEM−EXACT capping; *CI excludes 0; †GPQA: boxed judge.

		MATH-500		AIME 2025		GPQA-Diamond		LiveCodeBench	
intervention (cause)	format	cap.	acc.	cap.	acc.	cap.	acc.	cap.	pass@1
untouched student (levels, %)	EXACT	15.0	83.4	49.7	26.1	27.0	48.7	41.1	7.8
	NO-SYSTEM	31.6	82.8	67.2	24.4	55.1	46.5	67.9	13.0
	gap	$+16.6^*$	—	$+17.5^*$	—	$+28.0^*$	—	$+26.8^*$	—
two formats from step 0 (ii)†	gap	-1.1	-0.8	-2.5	$+1.4$	$+0.4$	-0.1	—	—
repair D, one epoch after SFT (ii)	gap	$+1.6$	-0.2	$+3.3$	$+0.3$	—	—	—	—
decode guard TERMINATOR (iii, ii)	EXACT	-15.0^*	$+4.6^*$	-49.7^*	$+2.2^*$	-27.0^*	$+6.4^*$	—	—
	NO-SYSTEM	-31.6^*	$+21.7^*$	-67.2^*	$+9.7^*$	-55.1^*	$+23.1^*$	—	—
science student, decode guard	NO-SYSTEM	-60.4^*	$+44.7^*$	-81.7^*	$+11.2^*$	-60.7^*	$+25.9^*$	—	—
budget forcing (s1) at the budget	EXACT	-15.0^*	$+6.2^*$	—	—	—	—	—	—
give-up: own states + replay†	EXACT	-6.7^*	-0.7	-38.3^*	-1.1	-20.5^*	-2.8	—	—
	NO-SYSTEM	-23.4^*	-0.6	-56.4^*	-1.9	-48.1^*	-2.9	—	—
give-up: teacher + replay, $2\times$ lr†	EXACT	-0.1	-1.7	-15.6^*	$+4.4$	-4.3^*	-3.4^*	-8.0^*	$+1.4$
	NO-SYSTEM	-16.3^*	-1.6	-28.1^*	$+5.6^*$	-33.1^*	-8.6^*	-41.1^*	-5.9^*
20 released checkpoints, TERMINATOR	own	-4.8^*	$+1.4^*$	—	—	—	—	—	—

After the fact, 86 reformatted rows repair in one epoch. Of four arms that continue the student on 3,420 rows (Figure 6), the three that format some rows under NO-SYSTEM or HISTORICAL repair and replay alone (C) does not; D, original examples partly reformatted, repairs without significant side effects, already at **86 reformatted rows, 0.3% of the training set**, and reproduces on a second seed, the 32B student and the science student (Appendix A16).

7 RELATED WORK

Why generation does not end. Non-termination was first studied as a property of decoding: incomplete decoding algorithms can produce infinite sequences (Welleck et al., 2020a), maximization-based decoding degenerates into repetition (Holtzman et al., 2020), and search in translation favors empty or short outputs (Stahlberg & Byrne, 2019; Murray & Chiang, 2018). We locate it in the fine-tuning data, in what accompanied the demonstrated stops. For reasoning models, Pipis et al. (2026) and Duan et al. (2026) explain loops by learning errors and self-reinforcement; ours are the greedy form of cause (iii), whose persistence outlasts them at temperature 1.0. Practitioners report students that never stop after fine-tuning with a changed end token (hxs91, 2024; axolotl-ai-cloud contributors, 2025), and distilled students over-generate on long traces (Li et al., 2025; Yu et al., 2025); cause (i) explains the first, and the identity shows which branch the excess fills.

How to stop, and how to measure it. Budget forcing (Muennighoff et al., 2025) and our decode guard force the commitment the principle requires at a fixed position; an earlier TERMINATOR (Nagle et al., 2026) and Dong et al. (2026) learn when to exit from features of the prefix, and Zhang et al. (2025a) read correctness from hidden states, which separate a failure from a slow success barely better than elapsed position (§5). Chen et al. (2024) count tokens past the first correct answer and Kaiser et al. (2026) factor success through completion; we split capping by correctness and show that computed accuracy hides the delivered loss. On-policy work (Yang et al., 2026; Luo et al., 2026; Fu et al., 2026) blames the objective; our causes come from what offline SFT data never showed. Lin et al. (2025) and format-sensitivity work (Sclar et al., 2024; Mizrahi et al., 2024; Jiang et al., 2025) vary wording and read accuracy; we vary task-free scaffolding and read termination.

8 CONCLUSION: DELIVERING AN ANSWER IS ITS OWN CAPABILITY

Computing an answer and delivering it are separate capabilities, and fine-tuning can keep the first while breaking the second. A student that will not stop fails one decision, closing the thought and the turn, in the three ways we isolate, because it stops only under conditions present at every stop it was shown: a supervised end-of-turn token, the system turn and a converged answer. TERMINATOR guards each with a data check, a second format and a decode guard that forces an answer at a fixed token position. A trained stop for cause (iii) would have to read what no demonstration carries, whether the student's answer has settled; in an exploratory test its forced answers expose it (Appendix A11).

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2503.04697.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.744/.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

axolotl-ai-cloud contributors. Sample packing masks the <end_of_turn> token. GitHub discussion, January 2025. axolotl-ai-cloud/axolotl discussion #2259, https://github.com/axolotl-ai-cloud/axolotl/discussions/2259.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. URL https://arxiv.org/abs/2405.14782.

Kang Chen, Fan Yu, Junjie Nian, Shihan Zhao, Zhuoka Feng, Zijun Yao, Heng Wang, Minshen Yu, and Yixin Cao. Thinking traps in long chain-of-thought: A measurable study and trap-aware adaptive restart. arXiv:2601.11940, January 2026.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.03883.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhe Dong, Fang Qin, and Manish Shah. When does learning to stop help? A cost-aware study of early exits in reasoning models, 2026. arXiv:2606.30852.

Wei Du, Shubham Toshniwal, Branislav Kisacanin, Sadegh Mahdavi, Ivan Moshkov, George Armstrong, Stephen Ge, Edgar Minasyan, Feng Chen, and Igor Gitman. Nemotron-Math: Efficient long-context distillation of mathematical reasoning from multi-mode supervision. *arXiv preprint arXiv:2512.15489*, 2025.

Zenghao Duan, Liang Pang, Zihao Wei, Wenbin Duan, Yuxin Tian, Shicheng Xu, Jingcheng Deng, Zhiyi Yin, and Xueqi Cheng. Circular reasoning: Understanding self-reinforcing loops in large reasoning models, 2026. arXiv:2601.05693.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.06514.

Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. Concise reasoning via reinforcement learning. arXiv:2504.05185, April 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes, 2026. arXiv:2603.25562.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2023. URL https://zenodo.org/records/10256836.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://www.nature.com/articles/s42256-020-00257-z.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

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

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026. Oral; arXiv:2506.04178.

Jinyi Han, Ying Huang, Ying Liao, Zishang Jiang, Xikun Lu, Haiquan Zhao, Xinyi Wang, Guanghao Zhou, Sihang Jiang, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. Your models have thought enough: Training large reasoning models to stop overthinking. arXiv:2509.23392, September 2025.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? Rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 19889–19899. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.1006/.

hxs91. [Bug]: The final reason why you will get a model that cannot stop generation when you finetune the Qwen2.5-7b-base use Lora and a non-<|endoftext|> token as eos_token. GitHub issue, November 2024. QwenLM/Qwen3 issue #1064, https://github.com/QwenLM/Qwen3/issues/1064.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Chatbug: A common vulnerability of aligned llms induced by chat templates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025. URL https://arxiv.org/abs/2406.12935.

Daniel Kaiser, Arnoldo Frigessi, Ali Ramezani-Kebrya, and Benjamin Ricaud. Beyond accuracy: Decomposing the reasoning efficiency of LLMs. arXiv:2602.09805, February 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. URL https://arxiv.org/abs/2210.15097.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. arXiv:2502.12143, February 2025.

Zhixuan Li, Jiangan Yuan, and Han Xu. Data and evaluation closed-loop for model capability enhancement, 2026. arXiv:2606.28471.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 2023.

Xiaofeng Lin, Hejian Sang, Zhipeng Wang, and Xuezhou Zhang. Debunk the myth of SFT generalization, 2025. arXiv:2510.00237.

Sanae Lotfi, Polina Kirichenko, Steven Li, and Zechun Liu. Quantized reasoning models think they need to think longer, but they do not, 2026. arXiv:2606.00206.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.556/.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying OPD: Length inflation and stabilization strategies for large language models, 2026. arXiv:2604.08527.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. arXiv:2402.18540.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. arXiv:2402.18540.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 22483–22512, 2024. URL https://openreview.net/forum?id=jzkpwcj200.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. URL https://arxiv.org/abs/2202.05262.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.759/.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. URL https://aclanthology.org/2024.tacl-1.52/.

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

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pp. 212–223. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-6322/.

Alliot Nagle, Jakhongir Saydaliev, Dhia Garbaya, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. TERMINATOR: Learning optimal exit points for early stopping in chain-of-thought reasoning, 2026. arXiv:2603.12529.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. The EOS decision and length extrapolation. In *BlackboxNLP Workshop at EMNLP*, 2020.

Charilaos Pipis, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. Wait, wait, wait... why do reasoning models loop? In *International Conference on Machine Learning (ICML)*, 2026. Spotlight; arXiv:2512.12895.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers)*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2305.14739.

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. URL https://openreview.net/forum?id=4FWAwZtd2n. arXiv:2408.03314.

Sangjun Song, Minjae Oh, Seungkyu Lee, Sungmin Jo, and Yohan Jo. ThinkBrake: Efficient reasoning via log-probability margin guided decoding. *arXiv preprint arXiv:2510.00546*, 2025. URL https://arxiv.org/abs/2510.00546.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1331/.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Ryan Teknium, Roger Jin, Jai Suphavadeeprasit, Dakota Mahan, Jeffrey Quesnelle, Joe Li, Chen Guang, Shannon Sands, and Karan Malhotra. Hermes 4 technical report. arXiv:2508.18255, August 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6287–6310. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-acl.375/.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, et al. Thoughts are all over the place: On the underthinking of o1-like LLMs. *arXiv preprint arXiv:2501.18585*, 2025.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.naacl-main.167/.

Chenxing Wei, Yao Shu, Mingwen Ou, Ying Tiffany He, and Fei Richard Yu. PAFT: Prompt-agnostic fine-tuning. *arXiv preprint arXiv:2502.12859*, 2025.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5553–5568. Association for Computational Linguistics, 2020a. URL https://aclanthology.org/2020.emnlp-main.448/.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020b.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

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

Yuxiao Yang, Tianrun Yu, Shangzhe Li, Kaixiang Zhao, Xuchao Zhang, Chetan Bansal, Huaxiu Yao, Taylor W. Killian, and Weitong Zhang. When EOS tokens disagree: Understanding length inflation in on-policy distillation, 2026. arXiv:2609.20511.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. LongProc: Benchmarking long-context language models on long procedural generation. In *Conference on Language Modeling (COLM)*, 2025a. URL https://arxiv.org/abs/2501.05414.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Conference on Language Modeling (COLM)*, 2025b. URL https://arxiv.org/abs/2502.03387.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. *arXiv preprint arXiv:2502.03373*, 2025.

Bin Yu, Hang Yuan, Haotian Li, Xueyin Xu, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen. Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models. arXiv:2505.03469, May 2025.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Conference on Language Modeling (COLM)*, 2025a. URL https://arxiv.org/abs/2504.05419.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Xuechen Zhang, Zijian Huang, Chenshun Ni, Ziyang Xiong, Jiasi Chen, and Samet Oymak. Making small language models efficient reasoners: Intervention, supervision, reinforcement. *arXiv preprint arXiv:2505.07961*, 2025b. URL https://arxiv.org/abs/2505.07961.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2311.10054.

Reproducibility statement

All generations are produced with fixed seeds and prompts passed as exact token ids; every table is regenerated from the raw generation files by a single script with a single answer judge and loop definition (Appendix A12.2). The byte-level formats of every prompt format for every base are given in Appendix A12.1. Training uses the public ms-swift trainer with ZeRO-3 and packing; the repair arms change only the data (Appendix A12.3). Every training arm and prospective test reported here was pre-registered in a timestamped file written before the corresponding job was created—predictions, endpoint and fallback included—and those files and the code accompany this submission as supplementary material; prompts, generated outputs and the repair datasets follow in the release.

The use of AI tools

In this work, we used generative AI tools to implement methods (AI coding assistants wrote much of the evaluation, probing, intervention and table-generation code and the training and evaluation job scripts, to the authors' specification and under their review), to clean and reformat data sets (the filtering scripts that build the science and code training sets), to design or provide feedback on the methodology and experiments (LLM-written adversarial reviews of drafts, several of whose suggested controls we ran, e.g. the closure-logit splice), to propose and refine hypotheses, and to interpret results. The model-generated reasoning traces used as training and repair data (§2, §6) are the object of study, not an assistive use. We have not used generative AI tools to generate other

synthetic data sets, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to translate, or to perform qualitative or thematic analysis. Additionally, we used generative AI tools to edit the paper for readability, to search for related work, to create figures and to format references. Every number in the paper is produced from the raw generation files by scripts in our code release, and the authors checked the reported numbers against those files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

APPENDIX CONTENTS

A1 Computed against delivered accuracy on released checkpoints 18
A2 Base format-awareness: a pre-registered null 18
A3 Scope and limitations 20
A4 A base-side quantity that screens initializations before training 20
A5 Restoring stop-after-wrong 22
A6 Cause (iii): where stopping was demonstrated, and what the failure state holds 22
A7 Where the format gap lands, and what each branch looks like 29
A8 Cause (i): the one-variable test 30
A9 Supporting tables for cause (ii) 31
A10 How SFT stacks guard the terminal token 35
A11 TERMINATOR: the decode guard in detail 36
A12 Protocol: prompts, metrics and training 41
A13 The effect, replicated and stress-tested 47
A14 Mechanism: where the change sits and what removes it 51
A15 Boundaries: what changes the size of the effect 57
A16 Repairs 64
A17 Additional material 67

WHAT IS IN THE APPENDIX

The appendix is organized so that a reader who doubts one specific claim can go to one place. The sections before §A12 hold the tables the body points to directly: computed against delivered accuracy (Table 5; §A1), the base format-awareness null (§A2), scope (§A3), the base-side screen (§A4), the stop-after-wrong and terminal-supervision ladders (§A5, Figure 8), the evidence for cause (iii) (§A6), the one-variable end-of-turn-token test (§A8, Tables 22 and 23), the arms of cause (ii) (§A9) and the decode guard in detail (§A11). **§A12** fixes what was actually run: the byte-level prompt for every format and every base, the outcome predicates and the judge, training data and decontamination, and evidence that our pipeline reproduces a standard harness on checkpoints previously scored with it. **§A13** is the evidence that the effect is not an artefact of one run: seven students (Table 46), an ordinary vLLM serving endpoint rather than our own runner (§A13.2), a larger token budget, an independently trained seed, and two further evaluation domains. **§A14** localizes it: the position-wise divergence probe (Table 28), the stop hazard by position, the prefix-swap rescue (Table 55), and every splice arm (Figure 9) including the sham, reverse (Table 24) and gated controls, its replications (Tables 62 and 63) and the 32B null (§A14.4). **§A15** is what changes its size: optimizer (Table 64), objective, data scale, and all 31 released checkpoints measured the same way. **§A16** is the repair: the dose curve down to 86 rows, replications on a second seed and the 32B student (Tables 75 and 76), and the out-of-domain arms (Table 26). **§A17** holds the full limitations, the extended related work and worked examples of the failure and its repair (Figure 7).

Table 5: **What is computed against what is delivered.** *computed* is the paper's judge and the harnesses': the last `\boxed{}` anywhere in the output. *delivered* counts it only when the thought was closed and the generation stopped on a stop token rather than the budget—the same equivalence check, a narrower eligible span; *delivered (stop only)* requires the natural stop but not the closed thought, the middle definition of §4. Qwen3-8B-Base student, MATH-500, n=500; Δ is the paired difference to EXACT, * significant. Under the lenient reading the perturbation looks free; under the strict one it costs 29.4 points, made of outcomes the paper already reports.

			computed accuracy %		**delivered accuracy** %		delivered (stop only) %	
prompt the harness sends	cap	closed	acc.	Δ [95% CI]	acc.	Δ [95% CI]	acc.	Δ [95% CI]
MATH-500, Qwen3-8B-Base CE student (seed 42), budget 8,192 (Table 42)								
EXACT (training format)	15.6	84.4	83.0	–	80.2	–	80.6	–
system + instruction line	13.8	86.4	84.6	+1.6 [−1.2, +4.4]	82.2	+2.0 [−1.0, +5.0]	82.4	+1.8 [−1.0, +4.8]
HISTORICAL	24.8	57.2	79.8	−3.2 [−6.4, +0.0]	51.4	−28.8* [−33.6, −24.4]	68.2	−12.4* [−16.4, −8.6]
NO-SYSTEM	33.0	54.6	82.4	−0.6 [−3.6, +2.4]	50.8	−29.4* [−34.0, −25.0]	62.8	−17.8* [−22.0, −13.8]
MATH-500 under four harness defaults, same student/seed/budget (Table 7)								
EXACT (training format)	15.6	84.4	83.0	–	80.2	–	80.6	–
simple-evals†	13.0	87.0	79.4	−3.6* [−6.8, −0.6]	78.8	−1.4 [−4.4, +1.6]	78.8	−1.8 [−4.8, +1.2]
lighteval†	19.6	75.2	77.8	−5.2* [−8.8, −1.6]	67.4	−12.8* [−17.0, −8.8]	72.0	−8.6* [−12.6, −4.8]
NeMo-Skills	25.4	57.6	81.2	−1.8 [−5.0, +1.4]	52.8	−27.4* [−32.4, −22.8]	67.8	−12.8* [−17.0, −9.0]
OpenCompass / EvalScope	26.4	67.6	80.6	−2.4 [−5.8, +1.0]	64.4	−15.8* [−20.0, −11.8]	70.4	−10.2* [−14.2, −6.4]

A1 COMPUTED AGAINST DELIVERED ACCURACY ON RELEASED CHECKPOINTS

Every chat-template cell of the first fifteen checkpoints of the harness panel (Appendix A12.4; the three added by the later census are not rescored; 15 checkpoints × four defaults, n=500), rescored with the paper's judge under both readings. Two cautions govern the table. simple-evals and lighteval ask for an `ANSWER:` line rather than a box, and released DeepSeek and Qwen3 checkpoints comply, so a box-only judge under-reads them there; each cell therefore also carries a *harness-native* reading (box or requested line, whichever is last, same equivalence check), and the format-clean comparison is the OpenCompass−NeMo-Skills pair, both of which request a box. And the part of a gap that is neither a capped nor an unclosed generation is extraction, not delivery—a correct answer after the closer in markup the anchored regexes miss—so only the capped and unclosed buckets are termination-attributable. Under the paper's judge the four single-dataset distills (our two students, OpenThinker-7B, TinyR1-32B-Preview) are the only rows whose delivered spread across defaults exceeds their computed spread (Table 6); all eight multi-format checkpoints sit within a point, and the larger harness-native gaps among them (Qwen3-14B under lighteval and simple-evals, 38.0 and 13.8 points) are extraction rather than delivery: 189 and 69 of their gap rows are closed thoughts with the requested answer line only inside the thought, and their capped and unclosed buckets hold at most one generation. On released checkpoints the gap is solved-then-capped almost exclusively—unclosed thoughts occur only in OpenThinker-7B—so released distills lose the end of turn rather than `</think>`, where our students lose both. The largest termination-attributable gap on a released checkpoint under an unmodified public default is TinyR1-32B-Preview under simple-evals: 460 of 500 correct by the harness's answer line, 308 delivered (30.4 points; 150 of the 152 gap rows solved-then-capped, one unclosed, one extraction); under the box judge and the format-clean NeMo-Skills default, 94.6 → 82.4 (12.2 points). Per-cell numerators, the extraction audit and the one-level-brace regex check (182 of 30,000 rows, at most 0.8 points on any cell) are in the release.

A2 BASE FORMAT-AWARENESS: A PRE-REGISTERED NULL

We pre-registered the conjecture that the family effect is really a property of the *base*: that Qwen base releases have already seen their own chat template in pretraining, so their students inherit a closing prior already conditioned on the system turn, while Llama and Mistral students must learn closing from scratch and learn it unconditionally. The prediction was that brittle-producing bases score high on base format-awareness and robust-producing ones low. The pre-registration was written before any generation existed and is included in the supplementary material.

Table 6: **Delivered accuracy varies with the harness where computed accuracy does not, and only for single-dataset distills.** Spread of computed and delivered accuracy across the four chat-template defaults under the paper's box judge and the harness-native reading, and the format-clean OpenCompass−NeMo-Skills difference; 95% CIs in brackets, * significant. Box-judge computed spreads for released rows are the answer-protocol artefact described above and are not quoted.

	spread, paper's judge		spread, native judge		OC−NS, paper's judge		
checkpoint	computed	**delivered**	computed	**delivered**	computed	**delivered**	class
our single-format students							
CE student, seed 42 *(ours)*	3.4 [−0.4, 7.2]	26.0* [21.4, 30.8]	3.4 [−0.4, 7.2]	26.6* [22.0, 31.4]	−0.6 [−4.0, +3.0]	+11.6* [+6.2, +17.2]	single-dataset
CE student, seed 123 *(ours)*	9.2* [4.8, 13.6]	40.2* [35.2, 45.0]	8.4* [4.2, 12.8]	40.8* [35.8, 45.4]	−1.2 [−5.0, +2.6]	+20.8* [+15.4, +26.4]	single-dataset
released checkpoints							
DS-R1-Distill-Qwen-7B	86.4* [83.0, 89.6]	86.6* [83.4, 89.8]	5.4* [2.2, 8.6]	5.4* [2.2, 8.6]	+1.2 [−1.4, +3.6]	+1.2 [−1.4, +3.6]	multi-format
OpenThinker-7B	52.8* [48.0, 57.6]	71.4* [67.2, 75.6]	32.4* [27.6, 37.4]	67.4* [62.8, 71.8]	+0.8 [−2.0, +3.6]	−0.4 [−3.4, +2.6]	single-dataset
OpenMath-Nemotron-7B	2.6* [0.8, 4.6]	2.4* [0.6, 4.2]	2.8* [1.0, 4.8]	2.4* [0.6, 4.2]	−1.4 [−3.2, +0.2]	−1.8 [−3.6, +0.0]	other
Qwen3-8B	66.0* [61.6, 70.4]	66.6* [62.2, 71.0]	1.2 [−1.4, 3.6]	7.8* [5.0, 10.6]	−1.0 [−2.6, +0.6]	−0.8 [−2.4, +0.8]	multi-format
DS-R1-0528-Qwen3-8B	91.0* [88.4, 93.6]	91.0* [88.4, 93.6]	7.6* [4.8, 10.6]	8.0* [5.2, 11.0]	−6.8* [−9.4, −4.4]	−7.0* [−9.6, −4.6]	multi-format
Light-R1-7B-DS	84.0* [80.6, 87.4]	83.8* [80.4, 87.4]	17.8* [14.0, 21.6]	17.4* [13.6, 21.2]	+0.8 [−1.2, +3.0]	+0.4 [−1.8, +2.6]	other
DS-R1-Distill-Qwen-14B	91.6* [89.2, 94.0]	91.6* [89.2, 94.0]	9.0* [5.6, 12.4]	9.0* [5.6, 12.4]	−0.6 [−2.8, +1.6]	−0.6 [−2.8, +1.6]	multi-format
DS-R1-Distill-Qwen-1.5B	69.6* [65.2, 73.8]	69.8* [65.4, 74.0]	7.0* [3.0, 11.0]	6.8* [2.8, 10.8]	+3.4* [+0.2, +6.6]	+3.4* [+0.2, +6.8]	multi-format
TinyR1-32B-Preview	88.2* [85.4, 91.0]	91.4* [88.8, 93.8]	7.6* [4.8, 10.4]	33.0* [28.4, 37.6]	+0.2 [−1.4, +1.8]	+12.2* [+9.0, +15.4]	single-dataset
DS-R1-Distill-Qwen-32B	92.0* [89.6, 94.2]	91.8* [89.4, 94.0]	9.8* [6.4, 13.2]	9.4* [6.0, 13.0]	+4.4* [+2.0, +7.0]	+4.0* [+1.6, +6.6]	multi-format
Qwen3-4B-Thinking-2507	89.0* [86.0, 91.8]	89.0* [86.2, 91.8]	1.4 [−1.0, 3.8]	2.2 [−0.4, 4.8]	+0.0 [−1.6, +1.8]	−0.4 [−2.0, +1.2]	multi-format
Qwen3-14B	83.4* [80.0, 86.8]	84.0* [80.8, 87.2]	6.6* [3.8, 9.4]	44.0* [39.4, 48.4]	−0.4 [−1.8, +1.0]	−0.2 [−1.6, +1.2]	multi-format
Light-R1-14B-DS	87.8* [84.8, 90.6]	88.2* [85.0, 91.0]	14.0* [10.4, 17.8]	14.2* [10.6, 18.0]	−0.8 [−2.8, +1.0]	−1.0 [−3.0, +1.0]	other

Table 7: **A termination number is a property of the pair (model, harness).** Eighteen checkpoints, one benchmark, one decoder, one judge, 500 generations per cell; within a row only the prompt bytes differ, each being what that harness sends for MATH-500 by default, re-derived from its pinned checkout. The last five rows were added under a pre-registration that named TinyR1-32B-Preview as the released checkpoint that would swing and simple-evals—the only default that sends a system message—as its harshest column, its parent as flat, and the other three as flat: all three predictions held. *spread* is max−min. Budgets differ by row (8,192 for our students, 16,384 for the released checkpoints), but the spread is not a budget artefact: re-capping every generation at a common 8,192 tokens leaves the picture unchanged (OpenThinker-7B 12.6, DeepSeek-R1-0528-Qwen3-8B 19.0, TinyR1-32B-Preview 34.0; column in the released report). Run-to-run noise is about one point. **The ordering is model-specific**: our students cap least under simple-evals, OpenThinker-7B and TinyR1-32B-Preview most under it—so a harness cannot be ranked once and the ranking reused, which is what makes the number belong to the pair rather than to either member. Accuracy is left out because two of these harnesses ask for a plain `ANSWER:` line rather than a boxed answer, which no single judge scores comparably. lm-evaluation-harness applies no chat template at all and sits on a different axis (Appendix A12.4).

	cap % under each harness default					
checkpoint	simple-evals	lighteval	NeMo-Skills	OpenCompass	**spread** [95% CI]	loop
our single-format students						
CE student, seed 42 *(ours)*	13.0	19.6	25.4	26.4	**13.4** [9.8, 17.2]	12.0
CE student, seed 123 *(ours)*	13.8	28.4	36.8	25.6	**23.0** [18.4, 27.6]	21.2
released checkpoints						
DS-R1-Distill-Qwen-7B	1.8	1.8	2.6	4.2	**2.4** [0.8, 4.0]	0.2
OpenThinker-7B	22.6	11.6	6.6	7.6	**16.0** [12.0, 20.0]	17.4
OpenMath-Nemotron-7B	0.8	2.6	2.0	1.8	**1.8** [0.6, 3.2]	0.0
Qwen3-8B	4.2	2.6	3.6	4.8	**2.2** [0.6, 4.0]	0.0
DS-R1-0528-Qwen3-8B	3.6	5.8	4.6	12.8	**9.2** [6.8, 11.8]	0.0
Light-R1-7B-DS	1.4	1.6	1.6	2.6	**1.2** [0.0, 2.4]	0.2
DS-R1-Distill-Qwen-14B	1.2	1.4	1.2	2.4	**1.2** [0.0, 2.6]	0.0
DS-R1-Distill-Qwen-1.5B	6.2	4.8	7.4	8.2	**3.4** [1.0, 5.8]	0.6
TinyR1-32B-Preview	33.2	1.0	14.0	2.0	**32.2** [28.0, 36.4]	35.4
DS-R1-Distill-Qwen-32B (its parent)	1.6	0.6	1.2	1.8	**1.2** [0.0, 2.4]	0.2
Qwen3-4B-Thinking-2507	2.8	6.0	7.0	7.4	**4.6** [2.6, 6.6]	0.0
Qwen3-14B	2.4	3.0	3.8	3.4	**1.4** [0.0, 2.8]	0.0
Light-R1-14B-DS	2.2	2.8	3.0	4.0	**1.8** [0.2, 3.4]	0.4
s1.1-32B	5.4	4.0	4.0	3.4	**2.0** [0.0, 4.0]	1.0
OpenReasoning-Nemotron-32B	6.4	5.8	6.4	5.8	**0.6** [−1.2, 2.4]	0.0
QwQ-32B	1.8	1.8	2.8	2.0	**1.0** [−0.4, 2.4]	0.0

The prediction is inverted, not merely unmet (Table 8). No fine-tuning; each pretrained base formatted through its own chat template on 100 MATH-500 problems at two seeds (n=200 per cell), decoded exactly as in §2, reusing the paper's own formatting and evaluation code unmodified.

The same run gives the negative control used in §4: formatted with and without the system turn, the two Qwen bases are insensitive (Δcap -0.5 and $+1.0$), while the two robust-producing bases

Table 8: **Base format-awareness runs opposite to the brittleness each base produced.** Rate at which an untuned base spontaneously emits its own end-of-turn token, n=200 per row. The two bases whose students break hardest never emit theirs, falling back on the document-level end-of-text token, although both ship a chat template declaring the end-of-turn token as the EOS token. The pre-registered separation was $+20$ points in the other direction; observed, -62.0. Student Δcap from Table 68.

pretrained base	its student	own end-of-turn token	doc-EOS	student Δcap
Qwen3-8B-Base	brittle	0.0% [0.0, 1.9]	98.0%	+17.4
Qwen2.5-7B	brittle	0.0% [0.0, 1.9]	99.5%	+36.4
Llama-3.1-8B	robust	0.0% [0.0, 1.9]	36.0%	+3.2
Mistral-7B-v0.3	robust	62.0% [55.1, 68.4]	—	+2.6
Llama-3.1-8B-Instruct	robust	94.5% [90.4, 96.9]	0.0%	+0.2

move in the opposite direction to their students (Llama-3.1-8B $64.0 \rightarrow 24.5\%$, Mistral-7B-v0.3 $38.0 \rightarrow 18.0\%$). The system-turn dependence is therefore manufactured by the fine-tuning rather than inherited. We report the awareness result only as a negative: an exploratory natural-stop variant separates the four primary bases but collapses once the instruct and 32B checkpoints are included, reproducing the failure mode of the base-prior predictor it was meant to replace.

A3 Scope and limitations

One teacher per domain, one trainer, one harness in depth. Our brittle students are all Qwen-derived (Qwen3-8B-Base, Qwen2.5-7B, DeepSeek-R1-Distill-Qwen-32B) and the robust ones Llama-3.1, Mistral and Qwen2.5-14B, with the Llama-based R1 distill short of the endpoint—fifteen runs over eight initializations, confounded with capability, with own-format capping, and with how much the deletion removes (Mistral's template has no system slot at all). Cause (ii) is localized, at the two closing tokens, on the 8B and 7B students only: the splice shows the closing conditional there is *sufficient* to restore termination, not which circuit computes it, and it is inert on the 32B student. The two-format and mirror arms (Appendix Table 25) are one base, one seed and one pair of formats each; the mirror arm's null says that *this* recipe does not produce the added-turn failure TinyR1 shows, not that no recipe does. The format fix is shown to survive the ladder's added traces on one training seed (Table 10); cause (i) was never crossed with the others, since the main recipe keeps the end-of-turn token. Cause (iii) is removed at decoding time, not in training: on one training seed, the best give-up training dose halves the wrong branch with no significant accuracy change but a 2.8-point GPQA-Diamond point estimate, stronger doses cost two to three points with replay and most accuracy without (Appendix A6),and none was crossed with a second seed or LiveCodeBench beyond one arm; a trained stop would need the value of continuing from the student's own state, a signal our demonstrations do not carry. Nucleus truncation does not create the format gap: with top-p disabled it is 16.4 points [12.4, 20.4] against 17.4 (one seed). The three causes are those we can localize in supervised, mathematics-trained students under fixed decoding and a single turn; budget exhaustion on genuinely hard problems, decoding-induced degeneration, serving and stop-set errors, tool loops and RL-induced length growth are outside this study. Cause (i)'s prevalence is measured on two stacks' defaults and two corpora (Table 31), not on released training pipelines. All measurements are single-turn, single-sample, no tool use, one temperature. The pre- versus post-fine-tuning contrast rests on five initializations. The 95% CIs resample problems within one training run and do not carry run-to-run training variance (Appendix A17.1).

A4 A base-side quantity that screens initializations before training

§4 reports the paper's cleanest discriminator of *which* closing token breaks: the shift the perturbation applies to a closing logit over the margin that logit holds, measured on the fine-tuned student. That is a description after the fact. Here we ask whether the same measurement on the *pretrained base*, before any fine-tuning, predicts whether the student will break at all.

Definition. For base M and closing token t, teacher-forced on the held-out teacher traces of Appendix A14.1 under the training format and under NO-SYSTEM, $D(M, t) =$

$\text{med margin}_{\text{EXACT}}(M,t) - \text{med margin}_{\text{NO-SYSTEM}}(M,t)$ in logits, with margin the closing token's logit minus its strongest competitor's; this is the numerator of the student-side ratio, computed by the same unmodified probe. The ratio itself is undefined on a base: no base ranks a chat closing token first at a closing position (margins -0.75 to -22), so the denominator is negative and the ratio changes sign. The primary statistic, fixed in a pre-registration written before any new measurement, is $D_{\text{atomic}}(M) = \max_t D(M,t)$ over the closing tokens that are a single token in M's vocabulary (`</think>` is one token for Qwen3 and the DeepSeek distills, three for Qwen2.5, Llama, Mistral and Yi; end-of-turn is one token everywhere). The pre-registration also fixed five alternative specifications, and we report them all.

Table 9: **Measured on the untuned base, the closing-logit displacement separates the initializations whose students broke from those whose students held—and it is not a family label.** D in logits, median over ≈ 180 held-out traces, with a 95% CI that resamples traces. *cohort 2* repeats D_{atomic} on a disjoint set of 200 traces, blind. *student* Δ*cap* is the paper's seed-42 single-format student trained from that base, MATH-500, EXACT $\rightarrow$ NO-SYSTEM, with its 95% CI (the cells of Appendix Table 49). The lower block are bases that had no student when the quantity was defined; the two marked *predicted* were then trained with the paper's recipe as a prospective test.

pretrained base	`</think>`	D(`</think>`)	D(end)	D_{atomic} [95% CI]	cohort 2	student Δcap	student
The six initializations the paper trained (outcomes known before the quantity was defined)							
Qwen3-8B-Base	1 token	+3.66	+0.88	+3.66 [+3.38, +3.94]	+3.78	$+16.6$ $[+13.2, +20.0]$	**brittle**
Qwen2.5-7B	3 tokens	+1.06	+2.90	+2.90 [+2.60, +3.16]	+2.95	$+40.0$ $[+36.3, +43.7]$	**brittle**
DS-R1-Distill-Qwen-32B	1 token	+1.97	+0.00	+1.97 [+1.72, +2.16]	+2.03	$+39.6$ $[+35.2, +44.0]$	**brittle**
Llama-3.1-8B (base)	3 tokens	+0.81	+0.85	+0.85 [+0.66, +1.09]	+0.91	$+0.0$ $[-3.0, +3.0]$	robust
Llama-3.1-8B-Instruct	3 tokens	+0.94	-0.19	-0.19 [-0.38, +0.06]	+0.00	$+0.2$ $[-3.4, +3.8]$	robust
Mistral-7B-v0.3	3 tokens	-1.00	+0.00	+0.00 [-0.12, +0.00]	-0.12	$+0.7$ $[-2.1, +3.5]$	robust
Bases with no student when the quantity was defined (outcome-blind)							
Qwen3-4B-Base	1 token	+2.65	+0.68	+2.65 [+2.06, +3.13]	–	–	no student
Qwen2.5-14B	3 tokens	-0.66	-1.00	-1.00 [-1.14, -0.89]	–	$+0.2$ $[-1.8, +2.2]$	*predicted not brittle*: **not brittle**
DS-R1-Distill-Llama-8B	1 token	+3.50	+0.00	+3.50 [+3.19, +3.88]	–	$+4.5$ $[+2.3, +6.7]$	*predicted brittle*: short of endpoint (miss)
Yi-1.5-9B	3 tokens	+1.25	-0.38	-0.38 [-0.69, +0.12]	–	–	no student

What separates, and what does not. On the six initializations the paper trained, D_{atomic} puts the three brittle-producing bases at 1.97, 2.90 and 3.66 and the three robust-producing ones at -0.19, 0.00 and 0.85: a gap of 1.1 logits, surviving the removal of any one base, with non-overlapping 95% CIs at the boundary. A blind replication on a disjoint cohort of traces, with all four pre-registered predictions committed in advance, reproduces it (2.03–3.78 against -0.12–0.91, the three brittle-producing bases in the same order). The token the quantity flags is the token the student later broke at: Qwen3-8B-Base at `</think>` (3.66 against 0.88 for end-of-turn; its student's ratio 0.90 against 0.39), Qwen2.5-7B at end-of-turn (2.90 against 1.06; its student breaks there, §4). Two of the five alternative specifications do not separate, and we say so: reading end-of-turn alone (the one token atomic in every vocabulary) does not, because Qwen3-8B-Base's displacement, and the 32B distill's, sits at `</think>`; and the competitor-free log-probability form does not, because the 32B distill—already a reasoning model—closes with $p \approx 1$ under both formats, so a log-probability difference saturates at zero where the margin still moves by two logits. With three brittle against three robust the exact one-sided permutation floor is $p = 0.05$ and six specifications were examined, so nothing here is a significance claim; the pre-registration's own verdict is *partial*. Two further bases were undefined: the probe returned no margin for InternLM2.5 (7B and 20B), and Nemotron-3-Nano did not fit its lane.

It is not a Qwen label. In the six-base sample the outcome is perfectly collinear with family, so any quantity separating Qwen from non-Qwen would "predict" it. Four bases with no student break that collinearity in both directions: Qwen2.5-14B scores -1.00, below every robust-producing base, while DeepSeek-R1-Distill-Llama-8B scores $+3.50$, above the lowest brittle-producing one; Qwen3-4B-Base (2.65) and Yi-1.5-9B (-0.38) fall where their tokenizers would suggest. A tokenizer reading—a one-token `</think>` implies a large displacement—is already inconsistent with Qwen2.5-7B, whose `</think>` is three tokens and whose student is brittle at end-of-turn.

Prospective test. We therefore pre-registered the two discordant cases as predictions (Qwen2.5-14B: *not* brittle, against every Qwen-derived student so far; DeepSeek-R1-Distill-Llama-8B: brittle, at `</think>`, against every Llama-derived student so far), fixed the endpoint of §2, and only then trained both with the paper's exact recipe. After three epochs the Qwen2.5-14B student is not brittle—cap 15.1 $\rightarrow$ 15.3, $+0.2$ $[-1.8, +2.2]$ (first decoding seed alone $+0.8$ $[-2.2, +3.8]$), every one of its 847 perturbed natural stops on `<|im_end|>`—and after three epochs the DeepSeek-

R1-Distill-Llama-8B student, predicted brittle, moves cap 18.0 → 22.5, +4.5 [+2.3, +6.7] (first decoding seed alone +2.8 [−0.2, +5.8]), the predicted direction but short of the endpoint: the first prediction is met and the second missed. Every number for this initialization uses the checkpoint's own `tokenizer.json`, which reproduces the training input ids; our evaluation runtime's default loader silently drops every space and newline for this vocabulary, and an earlier read-out through it gave +2.5 and $D = +2.50$ (the prediction was registered from that D; the corrected one points the same way). The untuned base itself moves +2.2 [+0.3, +4.1] under the same deletion. The read-outs are the last two rows of Table 9; the pre-registration file, with the fallback rule for a checkpoint short of three epochs, is in the supplementary material.

A5 Restoring stop-after-wrong

Correction: the arms restored almost no wrong endings. The pool a legacy string matcher labelled incorrect came from a string matcher that returns false whenever the normalized gold answer is shorter than two characters and rejects many equivalent free-form answers; the release itself keeps only solutions its own judge accepted. Re-judged against gold (integer comparison, then `math-verify`, then reading a random sample of the remaining rejects), about 3.4% [1.5, 7.6] of the pool is genuinely wrong (15 of 6,105 integer answers). The 5, 15 and 30% arms therefore made about 0.2, 0.5 and 1.0% of the training rows wrong-ending and otherwise added stops after correct answers. They remain informative as the case the principle predicts to fail: more stops shown after an answer do not reach a failure state that holds none (§5).

The arms as pre-registered: traces from that pool (each has exactly one closed thought and 53,685 of 53,694 carry a boxed answer; 53,681 eligible) replace 5, 15 or 30% of the 30,000 correct rows, total rows and every other setting byte-identical to the main student. Cells are three-epoch read-outs; differences are paired against the main student (* significant; 95% CIs in brackets). The identity cap = wrong × (1 − stop-after-wrong) + solved-then-capped holds on every cell by construction, up to rounding; its use here is as a bound, since a wide null on stop-after-wrong, multiplied by the wrong rate, is a narrow one on capping. The format gap column is NO-SYSTEM minus EXACT capping for the same student: on Qwen3 the wrong traces move it from +16.6 mostly within the single-format training-seed band of Table 30 (+20.8 at 15%; +23.6 and +32.5 at 30% on two training seeds), except +41.6 at 5% (the 5% arm loops on 35.9% of its perturbed generations), and leave it on Qwen2.5-7B (+40.0 → +39.4). In the last row the format fix survives the wrong-trace data: trained on both formats from the start, a student given the same 15% of wrong traces has a format gap of +0.9 (not significant), where the single-format student with those traces gives +20.8 (one training seed); cause (i) is not crossed here, since the main recipe keeps the end-of-turn token.

Table 10: **Traces labelled wrong-ending, about 97% of them correct, move neither the disposition measurably nor the capping.** Stop-after-wrong moves by a few points (not significant), own-format capping stays where it was, and in the last row the format fix survives the same data.

student (Qwen3-8B-Base unless noted)	format	cap	acc	wrong	stop-after-wrong	solved-then-capped	format gap
main student, 0% wrong traces	EXACT	15.0	83.4	16.6	23.5	2.3	+16.6*
5% wrong traces	EXACT	14.4 (−0.6)	82.9 (−0.5)	17.1	27.5 (+4.0 [−3.6, +11.6])	2.0	+41.6*
15% wrong traces	EXACT	15.8 (+0.8)	81.4 (−2.0)	18.6	31.2 (+7.7 [−0.9, +16.0])	3.0	+20.8*
30% wrong traces	EXACT	15.5 (+0.5)	82.4 (−1.0)	17.6	25.0 (+1.5 [−6.0, +9.0])	2.3	+23.6*
30% wrong traces, seed 123	EXACT	15.7 (+0.7)	82.8 (−0.6)	17.2	24.4 (+0.9 [−6.8, +8.5])	2.7	+32.5*
Qwen2.5-7B, 30% wrong traces	EXACT	20.7 (+0.8)	74.8 (−0.1)	–	27.8 (+0.3 [−6.4, +6.8])	2.5	+39.4* (sibling +40.0*)
both formats + 15% wrong (capstone)	EXACT	15.7 (+0.7)	82.8 (−0.6)	17.2	25.0 (+1.5 [−6.9, +9.6])	2.8	+0.9 [−1.1, +3.0]

A6 Cause (iii): where stopping was demonstrated, and what the failure state holds

Every stop the student was shown was made while holding a converged, correct answer; its failure state holds no settled answer, the student represents that state as failing and keeps going, and stops demonstrated in its own stuck states restore part of its stop-after-wrong rate. This appendix gives the evidence behind §5 in seven parts: the corpus (Table 12), the failure state (Tables 13 and 14), generality (Table 15), acquisition (Table 16), stops shown after an answer (Table 17), temperature (Table 18) and give-up sentences shown from answer-less states (Table 19). Unless noted, read-outs are the Qwen3-8B-Base main student on MATH-500 under its own format (EXACT) at an

Table 11: **The terminal-supervision ladder behind Table 3.** Every arm under the training format and with the system turn removed; brackets give the 95% CI of the paired difference from the untouched cross-entropy student, added to that student's value. The arm Appendix A15.1 calls stop up-weighting is the BalancedAdamW series (B) at ratio ≈66: two interventions at once; its accuracy collapse belongs to that implementation, and it goes below the cross-entropy control on off-format capping only by paying that accuracy. A plain cross-entropy weight of 16 halves off-format capping at −0.1 accuracy, and the plain-AdamW series (A) is not monotone in the ratio.

				EXACT			NO-SYSTEM	
series	ratio	seed	n	cap	acc	stop|wrong	cap	acc
base model	–	–	500	1.8	55.2	96.0	1.4	62.6
CE student	–	42	1000	15.0	83.4	23.5	31.6	82.8
A	1	42	1000	17.1 [15.1,19.2]	81.7 [79.8,83.5]	24.6	49.8 [45.8,53.9]	78.4 [76.0,80.9]
A (repl.)	1	42	1000	15.5 [13.4,17.7]	84.2 [82.0,86.3]	23.4	42.3 [38.8,45.8]	80.7 [78.3,83.1]
A	3	42	1000	17.4 [15.3,19.6]	81.9 [79.7,84.1]	24.3	36.2 [32.7,39.7]	79.8 [77.6,82.0]
A	10	42	1000	15.2 [13.2,17.3]	84.0 [81.9,85.9]	25.0	45.9 [42.3,49.7]	79.7 [77.3,82.1]
A	30	42	1000	15.2 [13.0,17.4]	83.4 [81.1,85.6]	27.7	27.1 [23.6,30.7]	76.8 [74.1,79.5]
A	100	42	1000	16.6 [14.6,18.7]	81.9 [79.9,83.8]	23.8	27.7 [24.0,31.4]	80.9 [78.6,83.2]
control	–	42	1000	15.2 [13.3,17.2]	82.7 [80.8,84.6]	25.4	16.0 [12.7,19.2]	82.7 [80.7,84.7]
B	0	42	1000	16.5 [14.5,18.6]	81.5 [79.2,83.7]	25.4	22.2 [18.6,25.8]	81.9 [79.7,84.2]
B	3	42	1000	15.4 [13.3,17.6]	83.5 [81.3,85.6]	28.5	11.0 [7.7,14.4]	70.4 [67.7,73.1]
B	3	123	1000	16.5 [14.4,18.7]	81.8 [79.4,84.0]	26.9	12.1 [8.8,15.3]	66.2 [63.5,69.0]
B	∼66	42	1000	13.8 [11.8,15.8]	83.6 [81.8,85.4]	28.7	11.1 [7.8,14.3]	62.3 [59.5,65.1]
B	∼66	123	1000	16.5 [14.5,18.5]	82.3 [80.4,84.2]	25.4	10.7 [7.4,14.0]	65.5 [62.7,68.3]
B	∼66	7	1000	15.7 [13.8,17.6]	82.7 [80.7,84.7]	23.7	14.4 [11.0,17.8]	71.0 [68.1,73.8]

8,192-token budget; every experiment on a model was pre-registered before its jobs ran (files in the supplementary material), and every table is generated from the run receipts.

Every demonstration stops holding a correct answer. Every one of the 30,000 teacher traces ends on the correct answer, 99.99% state it inside the thought, and none is unfinished. The teacher commits early and re-checks: its first candidate sits at a median 53% of the thought, 94.5% of traces re-derive after it, and a trace states its answer a median 10 times before closing a median 45 tokens after the last statement (Table 12). Three filters produce this. The release's correctness judge removed every failed generation, 21.5% of the teacher's generations on these very problems, so wrong and unfinished attempts never reached the release; our label check removed 53,694 further rows, about 3.4% of them genuinely wrong (Appendix A5); our length cut kept 30,000 of 46,306 verified rows and discarded the long searches (median thought 32,282 tokens, 59.1% longer than 30,000). The filtered corpus inverts the relation between difficulty and length: on the problems the teacher solves at most a quarter of the time, the kept generations end by 28,475 tokens and none passes 30,000, while the same bin before filtering has a median of 10,499 and 22% past 30,000; across the corpus the teacher's pass rate correlates positively with thought length (Spearman 0.23, against 0.05 before filtering). The corpus therefore holds no long unconverged search, and no stop from one.

A larger budget leaves most stuck generations stuck. Continuing the same own-format samples from 8,192 to 30,000 tokens, 60.0% of the 75 wrong-and-capped generations are still capped without a correct answer, 22.7% go on to stop correct and 14.7% stop wrong; stop-after-wrong rises from 22.4 to 38.4%, against the base's 95.7% at 8,192 (Tables 12 and 16).

The stuck state holds no settled answer, and the state sets its commit probability whichever model reads it. We teacher-force the 127 own-format wrong-and-capped generations (both decoding seeds) and four comparison groups, cut at the last line break before 25, 50, 75 and 100% of the trajectory (Table 13). The probability of starting a commit phrase stays at most 0.02 at every cut of the stuck trajectories, and only 1 of the 127 has any line break with $p(\texttt{</think>}) > 0.1$; solved generations reach 0.996 at their close (difference −0.98, significant). The base reading the same stuck text gives the same commit probability (difference −0.0001, not significant) and the same forced-answer accuracy (29.1 against 26.8%, not significant). Forced to answer at the end, the student is right 26.8% of the time, near the accuracy of its own resamples on these problems

Table 12: **Every demonstration stops holding a correct answer, a median 45 tokens after last stating it; the searches that did not converge were removed by three filters.** Nemotron-Math-v2 teacher traces as trained on (Qwen3 tokens). The last block continues the main student's own-format MATH-500 generations from 8,192 to 30,000 tokens (same samples, 500 problems).

quantity	value
What every demonstration holds when it stops (30,000 traces)	
trace states an answer inside the thought	99.99%
trace left unfinished	0
answer statements per trace, median	10
first candidate, median position in the thought	53%
re-derivation after the first candidate	94.5%
tokens from the last answer statement to `</think>`, median [p25, p75]	45 [19, 111]
What the three filters removed	
(1) release judge: teacher generations on these problems that failed	21.5%
(2) our label check: rows removed (genuinely wrong)	53,694 (3.4%)
(3) our length cut: verified rows kept	30,000 of 46,306
cut rows: median thought tokens; share $>$ 30,000 tokens	32,282; 59.1%
Lengths, and what a larger budget buys the student	
thought tokens, median / p90 / max	7,827 / 16,601 / 37,731
hardest problems (teacher pass rate ≤ 0.25, $n = 3{,}066$): generation p99 / max	21,622 / 28,475
same bin before filtering ($n = 13{,}851$): median; share $>$ 30,000	10,499; 22%
student, own format, the 75 wrong-and-capped at 8,192, continued to 30,000:	
still capped, no correct answer / stop correct / stop wrong	60.0 / 22.7 / 14.7%
stop-after-wrong at 8,192 $\rightarrow$ 30,000 tokens	22.4 $\rightarrow$ 38.4

(22.3%); its forced answers change across the four cuts (mean 2.2 distinct of 4) and carry low probability (0.34 at the 75% cut, against 0.70 for solved and 0.65 for wrong-then-stopped generations). Only 44.9% of stuck texts write any candidate answer, 15.0% in a box, and the correct answer is among the candidates in 18.1%. Where a candidate is written, the commit probability rises fifteenfold (0.18 $\rightarrow$ 2.65%) and stays below that of solved generations (8.6%; difference -5.9 points, significant), so the missing candidate alone does not account for the state.

Only a settled answer closes the thought; writing one into the stuck state leaves it open. Appending "Thus final answer: `\boxed{F}`." with the student's own forced answer F leaves the median $p($`</think>`$)$ at 0.001 in stuck states, against 0.993 in solved and 0.994 in wrong-then-stopped ones. Solved-then-capped generations are as long as the stuck ones and close at 0.979, so context length leaves the close intact; within them the close follows the candidate's settledness (mean 0.80 when the forced answer is settled, 0.28 when it is not). Substituting the correct answer for the student's own lowers the close (wrong-then-stopped generations: median 0.99 $\rightarrow$ 0.25). The close is conditioned on a settled answer, which is what every demonstration held and what the failure state lacks.

The student represents the failing state as well as its base does, and the text alone carries it. Logistic probes on the residual stream after 1,024–4,096 generated tokens separate wrong-and-capped from solved generations at AUC 0.75–0.84 (layer 18); the base reading the same tokens reaches 0.69–0.84, and a TF-IDF probe on the text alone 0.82–0.88 (Table 14). Across 84 paired cells (layers, positions, contrasts) the student-minus-base median difference is 0.001, with 8 cells significantly negative and 4 positive. At the designated cell (layer 27, 2,048 tokens) the student scores 0.684 and the base 0.721 (-0.04 $[-0.07, -0.01]$); the gap does not replicate on a second training seed (-0.03 $[-0.07, +0.02]$) or without the system turn (-0.01 $[-0.04, +0.02]$). Asked "Am I confident my approach will reach a correct answer?" at the midpoint of its generation, the student ranks its failures at AUC 0.90 (the base on the same text 0.87), though yes and no together carry only 0.8% of its next-token probability there. The controls show what is read: against solved generations that are still running past 4,096 tokens the AUC falls to 0.56–0.59, and with both sides still lacking a boxed answer to 0.57–0.72. The student keeps re-deriving from a state it represents as failing; the knowledge is present and unused.

Table 13: **The stuck state holds no settled answer: its commit probability stays near zero whoever reads it, its forced answer is right near the student's resample rate, and a written answer does not close it.** Main student, MATH-500, EXACT, both decoding seeds; teacher forcing on existing generations cut at the last line break before each fraction of the trajectory (up to the first </think>). Commit probability: total probability of a line-start commit phrase (</think>, \boxed, "The final answer", ...). Forced answer: append "**Final answer:** \boxed{" and decode greedily; settledness is the greedy answer's sequence probability. Last row: F is the student's forced answer at the 100% cut. "(same)": identical text to the first column.

	student reads its own generation				base reads
generation	wrong-and-capped	solved-then-capped	solved, stopped	wrong, stopped	wrong-and-capped
generations (problems)	127 (81)	23 (23)	200 (178)	39 (30)	127 (81)
commit probability at 50% of the trajectory	0.015	0.067	0.062	0.035	0.015
commit probability at 75% of the trajectory	0.003	0.064	0.091	0.086	0.015
commit probability at 100% of the trajectory	0.019	0.126	0.996	0.996	0.019
forced answer correct at 100%	26.8	95.7	98.0	30.8	29.1
settledness of the forced answer at 75%	0.34	–	0.70	0.65	0.41
student's resample accuracy, same problems	22.3	–	96.0	31.6	(same)
text writes a candidate answer	44.9	100.0	100.0	100.0	(same)
in a box	15.0	100.0	99.5	94.9	(same)
correct answer among them	18.1	100.0	99.5	25.6	(same)
p(</think>) once the answer F is written, median	0.001	0.979	0.993	0.994	–

Table 14: **The failing state is linearly readable, as well by the base as by the student and as well from the text alone.** AUC of failing (wrong-and-capped) against solved-and-stopped generations at 1,024, 2,048 and 4,096 generated tokens (generations still running at that length); main student's MATH-500 EXACT generations, both decoding seeds. Probes: standardised L2 logistic regression, 5-fold cross-validation grouped by problem; the base reads the student's tokens. The two controls remove what the state shows (tokens still to go; whether an answer is held).

read-out (AUC, failing vs solved)	1,024	2,048	4,096
student residual stream, layer 18	0.84	0.79	0.75
base residual stream on the same text, layer 18	0.84	0.77	0.69
student residual stream, layer 27	0.79	0.68	0.62
base residual stream on the same text, layer 27	0.80	0.72	0.66
text only (TF-IDF of the prefix)	0.88	0.86	0.82
student asked "Am I confident ...?"	0.80	0.76	0.79
base asked the same on the same text	0.75	0.74	0.74
control: against solved generations still running past 4,096 (layer 18)	0.59	0.56	–
control: both sides still without a boxed answer (layer 18)	0.72	0.59	0.57
failing / solved generations	127 / 472	127 / 225	127 / 77

The signature is general. In 8 of 10 base–student pairs stop-after-wrong drops significantly (Table 15; the two exceptions are the science student and the Llama-3.1-8B pair, whose base writes no answer box). In all eight, 76–92% of the student's wrong-and-capped generations write no answer, 2–13% loop and at most 2% close the thought. The position-matched ratio of the stop hazard when wrong to that when solved is 0.08–0.16 for seven students (0.49 for the code student), against 0.81–1.00 for the other bases and 0.32–0.40 for the two bases that are themselves distilled reasoners. Among 21 testable released long-CoT checkpoints, 17 show the signature (ratio 0.06–0.31, against 0.50–0.92 for their measured parents); the four others miss only on the closed-thought share (12–24%, 13–21 generations). Both students with the teacher-forced wrong-close probe close after the teacher's wrong endings exactly as after its correct ones (ratio 1.00 under the own format). The stated criterion also required a late-budget hazard contrast, which holds in 5 of the 8 pairs; the whole-budget hazard ratio was added after the late comparator proved survivor-selected (only looping base generations survive past 2,000 tokens).

Table 15: **In every pair whose stop-after-wrong drops, the student's wrong-and-capped generations write no answer and the student stops almost only once it has solved.** MATH-500 under each student's training format, budget 8,192, common problems; maths traces unless noted, blind: trained after its outcome was predicted (Appendix A4). Δ stop-after-wrong: student minus base, 95% CI, * significant. Wrong-and-capped: student / base, % of generations. Behaviour columns: % of the student's wrong-and-capped generations. Stop hazard: Mantel–Haenszel position-matched hazard of stopping when wrong over when solved (1: stops regardless of correctness); Mistral's base solves too little to estimate it. Last row: released checkpoints at a 16,384 budget, ranges over the 17 that show the signature and over their measured parents (pretrained or instruction-tuned). All ten pairs: Table 21.

		wrong-and-	student's wrong-and-capped			wrong:solved hazard	
initialization	Δ stop-after-wrong	capped s / b	no box	loop	closed	student	base
Qwen3-8B-Base (main student)	-72.2^{*} $[-80.1, -63.9]$	12.7 / 2.0	85	8	1	0.09	0.81
Qwen3-8B-Base, code traces	-43.2^{*} $[-49.9, -36.4]$	23.0 / 2.0	89	2	1	0.49	0.81
Qwen2.5-7B	-69.0^{*} $[-85.4, -51.0]$	15.0 / 0.5	83	13	0	0.08	1.00
Qwen2.5-14B (blind)	-70.0^{*} $[-77.6, -62.0]$	12.3 / 1.5	92	3	1	0.11	0.92
DS-R1-Distill-Qwen-32B	-31.7^{*} $[-46.9, -16.1]$	5.2 / 5.0	92	8	0	0.16	0.40
DS-R1-Distill-Llama-8B (blind)	-29.2^{*} $[-37.8, -20.6]$	14.1 / 9.9	87	3	2	0.10	0.32
Llama-3.1-8B-Instruct	-59.9^{*} $[-75.8, -43.7]$	21.0 / 5.5	76	5	0	0.14	0.94
Mistral-7B-v0.3	-37.6^{*} $[-49.3, -24.9]$	34.0 / 38.0	81	3	2	0.11	–
released long-CoT checkpoints: signature in 17 of 21 testable						0.06–0.31	0.50–0.92

Training removes stopping, then restores it only where an answer is held. The base stops after 95.7% of its wrong answers and after every correct one. After 0.05 epoch the student stops on 22.6% of correct and 1.3% of wrong generations, after 0.10 epoch on almost none (1.1% and 0.0%); by 0.15 epoch stopping returns for 63.1% of correct generations and 20.7% of wrong ones, a one-epoch schedule ends at 94.7 and 33.2%, and three epochs end at 97.2 and 23.5% (the same 23.5% at 1.75 epochs; the one-epoch schedule's end is 9.7 points higher, significant). What returns is conditioned on the answer: holding a boxed wrong answer the three-epoch student stops 66.1% of the time, without one 1.8%, where the base stops 98.1 and 93.8% (Table 16). A student trained on the same problems with answer-only traces keeps the base's stop-after-wrong (99.7%), and 90.7% of its wrong generations hold a box.

Table 16: **Early training removes stopping altogether; what returns is stopping once an answer is held.** Main recipe (Qwen3-8B-Base, 30,000 maths traces), MATH-500, EXACT; steps 47–141 are the main run's early checkpoints; stop-after-wrong with 95% CI; the last three columns split the wrong generations by whether they hold a boxed answer. The answer-only student sees the same problems with answer traces and no thought.

			stop given		wrong with	stop given wrong and	
checkpoint	acc	cap	correct	stop-after-wrong	a box	a box	no box
base (step 0)	53.2	2.0	100.0	95.7 [92.9, 98.2]	44.9	98.1	93.8
step 47 (0.05 epoch)	16.4	95.2	22.6	1.3 [0.6, 2.2]	2.8	30.4	0.5
step 94 (0.10 epoch)	61.7	99.3	1.1	0.0 [0.0, 0.0]	–	–	–
step 141 (0.15 epoch)	72.0	48.8	63.1	20.7 [15.5, 25.9]	55.0	35.1	3.2
one-epoch schedule, end	77.4	19.2	94.7	33.2 [26.5, 40.1]	43.4	71.4	3.9
main, 1.75 epochs	83.0	15.4	97.1	23.5 [16.5, 30.9]	30.6	71.2	2.5
main, 3 epochs (the student)	83.4	15.0	97.2	23.5 [16.1, 31.8]	33.7	66.1	1.8
at a 30,000-token budget	85.4	10.0	98.8	38.4 [27.6, 50.0]	47.9	74.3	5.3
answer-only traces, 3 epochs	67.7	0.1	100.0	99.7 [99.0, 100.0]	90.7	100.0	96.7

Stops shown after an answer are learned as stops after an answer. The give-up student's corpus replaces 3,861 rows (12.9%) with the teacher's thought cut after its first boxed answer (median 80% of the thought) and followed by a fixed give-up sentence, a close and a hedged boxed answer; every other row is identical to the main corpus. Under the training format nothing moves (capping 14.9 against 15.0, stop-after-wrong +3.7 [−4.1, +11.6]; AIME25 alike), the sentence appears in 1 of 1,000 generations, a correct one, and all 46 stops after a wrong answer are confident wrong boxes (Table 17). Teacher-forced, the sentence is ten times likelier when a boxed answer is on the page ($\log_{10}$ ratio 1.00 [0.86, 1.14] over 6,621 states), on the teacher's traces (0.90) but not on the student's

own wrong generations (0.13 [−0.55, 0.64]). Without the system turn it becomes a looping exit: it appears in 265 generations, 245 of them correct and 219 capped; at the points where a thought closed (its own correct generations and teacher traces) its opening outranks </think> in half the cases, and in none under EXACT; loops rise from 16.3 to 45.3% and the format gap from 16.6 to 48.8 (+32.2 [+27.5, +36.9]), at −3.7 accuracy.

Table 17: **Give-up sentences shown after an answer never reach a failure under the training format and become a looping exit without the system turn.** Give-up student: 12.9% of rows cut after the teacher's first boxed answer and followed by a fixed give-up sentence; all other rows identical to the main student's. MATH-500 unless noted, both decoding seeds (AIME25: 240 generations); give-up sentence: % of generations containing it; Δ: paired difference from the main student, 95% CI, * significant. Probe: teacher forcing at line breaks, trace-bootstrap 95% CI.

student	format	cap	acc	stop-after-wrong	wrong-and-capped	loop	give-up sentence	Δ stop-after-wrong
main student	EXACT	15.0	83.4	23.5	12.7	1.3	0.0	
give-up student	EXACT	14.9	83.1	27.2	12.3	0.8	0.1	+3.7 [−4.1, +11.6]
main student	NO-SYSTEM	31.6	82.8	23.3	13.2	16.3	0.0	
give-up student	NO-SYSTEM	63.7	79.1	14.8	17.8	45.3	26.5	−8.4* [−15.6, −1.6]
main student	EXACT, AIME25	47.9	27.9	35.3	46.7	11.2	0.0	
give-up student	EXACT, AIME25	48.8	25.0	35.6	48.3	10.8	0.4	+0.3 [−7.4, +8.0]

Teacher-forced: $\log_{10}$ *of* p(*give-up sentence*) *with a boxed answer on the page over without (median difference)*

all 6,621 states (393 traces)	1.00 [0.86, 1.14]
teacher's traces	0.90 [0.71, 1.08]
the student's own wrong generations	0.13 [−0.55, 0.64]

Temperature removes the verbatim loops and keeps the persistence. Under greedy decoding the wrong-and-capped branch is 21.6% of generations and 81% of it is verbatim looping by the criterion of Pipis et al. (2026); at temperature 1.0 no loop remains and the branch is still 15.8% (Table 18). Of the 87 problems that loop greedily, 43 become non-looping wrong-and-capped at 1.0 and 5 become stops after a wrong answer. Stop-after-wrong stays between 10.7 and 22.4% at every temperature against the base's 90.8–98.0%, and the student's wrong branch exceeds the base's by 11.2–17.0 points at every temperature. On the 66 problems the student leaves wrong-and-capped at 0.6, the base stops with a wrong answer on 50 and solves 14.

Table 18: **Temperature removes the verbatim loops and leaves the persistence: stop-after-wrong stays below a quarter of the base's at every temperature.** Main student and its base, MATH-500, EXACT, 500 problems, one decoding seed. Verbatim loops: a 30-gram of token ids repeated at least 20 times, as a share of wrong-and-capped generations. Last column: paired item bootstrap, 95% CI, * significant.

	student					base		
temperature	cap	acc	wrong-and-capped	verbatim loops in it	stop-after-wrong	wrong-and-capped	stop-after-wrong	wrong-and-capped, student − base
0.0	38.4	75.8	21.6	81%	10.7	4.6	90.8	+17.0* [+13.2, +21.0]
0.3	21.2	81.2	15.0	51%	20.2	3.0	92.8	+12.0* [+8.6, +15.4]
0.6	15.6	83.0	13.2	9%	22.4	2.0	95.7	+11.2* [+8.0, +14.4]
1.0	20.2	80.6	15.8	0%	18.6	1.0	98.0	+14.8* [+11.6, +18.0]

Give-up training trades stop-after-wrong for accuracy. Each arm repairs the trained student with give-up rows: a trace cut inside the thought at a paragraph boundary before any candidate answer, followed by "I have not been able to finish this. My best answer is \boxed{X}.", the close and a terse "**Answer:** \boxed{X}" line, with X the last intermediate result and loss on that segment only. The teacher's states come from problems outside the corpus and MATH-500 (40–80% of the thought); the student's own come from its stuck generations on the MATH training split (11.6% of 14,772 rollouts, cut at 60–90%). Trained on 800 own-state rows alone for one epoch, the student stops after 40.2% of its wrong answers against 23.5% (+16.7 [+9.4, +24.2]), the wrong branch falls by 2.6 points and capping by 3.6, both significant, and accuracy stays put (−0.3); the arm meets its stated criterion on one training seed (Table 19). It never writes the give-up sentence: it stops with a boxed answer, and 28.8% of its stops end in the segment's terse answer line (0% for the untouched student). With 800 replay rows added, the same states give +6.7 [−1.7, +14.7]. From the teacher's answer-less states no one-epoch arm moves stop-after-wrong significantly (+3.4 with replay, +7.7 without, +4.7 on the Qwen2.5-7B student) and none uses the give-up sentence; on

the Qwen2.5-7B student the give-up sentence at answer-less states rose from a probability of about 10^{-22} to about 10^{-8}, by the same 31 nats on training states, held-out teacher states, the student's own stuck states and its correct ones, so the dose raised the sentence everywhere and selected no state. The teacher-state arm without replay still lowers the wrong branch (-3.0, significant): of the untouched student's 127 stuck rows, 49 now stop (36 correct) against 26 (19) when the untouched student is re-sampled. Every one-epoch give-up arm also shrinks the format gap, by 8.8–11.7 points, through fewer solved-then-capped generations under NO-SYSTEM (the sham by 3.0, not significant), and every higher dose all but closes it (-1.1 to $+0.7$; one collapsed stop-only arm reaches -15.0). A matched sham, the same 800 own stuck prefixes supervised on the student's own continuation, moves nothing ($+5.2$ $[-2.9, +13.0]$), so the own-state arm's gain over it ($+11.6$ $[+3.2, +20.4]$, wrong branch -2.6 $[-4.5, -0.7]$) comes from the demonstrated stop. Higher doses split by whether continuation is also supervised. Stop-only arms collapse at every higher dose and source: they stop on almost every problem (accuracy 4.6–7.1%; the first teacher-state arm closes the thought at its first token in all 1,000 generations), and so does a distilled forced-commit arm (29.8%). With replay, stop-after-wrong rises with dose and so does the cost: 2,000 teacher cuts over three epochs restore stop-after-wrong to 90.9% and cut the wrong branch to 1.8 at -3.2 points of accuracy, with 63 of 159 give-up stops on problems the untouched student solves, and the same data on Qwen2.5-7B cut its wrong branch from 18.2 to 0.8 at -2.0. The strongest own-state dose that keeps MATH-500 accuracy (-0.7) cuts capping to 8.3% on MATH-500, 11.4% on AIME 2025 (from 49.7) and 6.6% on GPQA-Diamond (from 27.0), but costs 2.8 points of GPQA accuracy $[-5.8, +0.3]$. A log-probability probe on the 2,000-cut arm shows why: the give-up sentence reaches $p{\approx}0.06$ per paragraph in the student's failures and 0.03 on paths to a solution, a ratio the base already had. Dose-ladder arms, marked by learning rate, use the same data files, recipe and read-out; they were pre-registered separately, later rungs were added before earlier results were read, and all are reported.

Table 19: **Give-up training trades stop-after-wrong for accuracy: stop-only arms collapse, low doses move little of the wrong branch, and every dose that restores stop-after-wrong costs accuracy.** Repairs from the trained student; give-up rows cut a trace inside the thought before any candidate answer and append a give-up sentence, the close and a boxed best result (loss on that segment only); replay rows are ordinary training rows; one-epoch arms at learning rate $2{\times}10^{-6}$, ≈2,000-row arms at 10^{-5}, ladder arms as marked. Paired against the untouched student, MATH-500, EXACT, both decoding seeds, 95% CI, * significant. Stuck rows now stopping: of the untouched student's 127 wrong-and-capped (problem, seed) pairs, those on which the arm stops (correct in brackets); the re-sampled untouched student gives the chance level. Format gap: NO-SYSTEM minus EXACT capping, * change significant. –: not computed.

arm (give-up rows)	Δ stop-after-wrong	Δ wrong-and-capped	Δ acc	give-up used	stuck rows now stop	format gap, untouched → arm
Teacher's answer-less states (Qwen3-8B-Base student)						
800 + 800 replay, 1 epoch	+3.4 [−4.1, +10.8]	+0.1 [−1.9, +2.1]	−0.9	0 / 1,000	38 (32)	+16.6 → +5.4*
800 alone, 1 epoch	+7.7 [−0.4, +16.1]	−3.0* [−5.0, −1.0]	+2.5*	0 / 1,000	49 (36)	+16.6 → +5.8*
≈2,000 + replay, 3 epochs	+67.4* [+59.0, +75.3]	−10.9* [−13.5, −8.4]	−3.2*	159 / 1,000	113 (24)	+16.6 → +0.7*
untouched student, re-sampled					26 (19)	
Teacher's answer-less states (Qwen2.5-7B student)						
800 + 800 replay, 1 epoch	+4.7 [−2.1, +11.6]	−0.5 [−2.6, +1.6]	−1.0	0 / 1,000	–	+40.0 → +29.3*
≈2,000 + replay, 3 epochs	+69.6* [+62.5, +76.3]	−17.4* [−20.4, −14.4]	−2.0	230 / 1,000	–	+40.0 → −0.1*
Teacher's answer-less states, dose ladder (Qwen3-8B-Base student)						
800 + 800 replay, 2 epochs, 10^{-5}	+16.5* [+8.0, +25.1]	−1.6 [−3.7, +0.5]	−1.9	36 / 1,000	–	+16.6 → −0.7*
800 + 800 replay, 2 epochs, $2{\times}10^{-5}$	+15.3* [+7.2, +23.8]	−1.5 [−3.6, +0.6]	−1.7	33 / 1,000	–	+16.6 → +0.4*
800 + 800 replay, 3 epochs, $3{\times}10^{-5}$	+5.5 [−2.5, +13.4]	+0.3 [−1.7, +2.3]	−1.7	16 / 1,000	–	+16.6 → −0.2*
800 alone, 3 epochs, 10^{-5}	+61.3* [+52.8, +69.2]	+1.4 [−2.1, +5.1]	−76.3*	711 / 1,000	–	+16.6 → −15.0*
800 alone, 3 epochs, $3{\times}10^{-5}$	+76.4* [+68.2, +83.9]	−12.6* [−15.3, −10.0]	−78.8*	77 / 1,000	–	+16.6 → +0.1*
The student's own stuck states (Qwen3-8B-Base student)						
800 + 800 replay, 1 epoch	+6.7 [−1.7, +14.7]	−0.2 [−2.4, +2.0]	−1.3	0 / 1,000	–	+16.6 → +4.9*
800 alone, 1 epoch	+16.7* [+9.4, +24.2]	−2.6* [−4.4, −0.9]	−0.3	0 / 1,000	–	+16.6 → +7.8*
sham: same prefixes, own continuation	+5.2 [−2.9, +13.0]	+0.0 [−2.1, +2.0]	−1.2	0 / 1,000	–	+16.6 → +13.6
800 + 800 replay, 2 epochs, 10^{-5}	+37.2* [+27.2, +47.0]	−5.9* [−8.1, −3.7]	−0.7	58 / 1,000	–	+16.6 → −0.1*
800 + 800 replay, 2 epochs, $2{\times}10^{-5}$	+36.3* [+26.9, +45.6]	−4.7* [−6.9, −2.5]	−3.3*	81 / 1,000	–	+16.6 → −1.1*
800 + 800 replay, 3 epochs, $3{\times}10^{-5}$	+60.2* [+51.3, +68.8]	−9.4* [−11.8, −7.1]	−3.7*	148 / 1,000	–	+16.6 → +0.4*
800 alone, 3 epochs, 10^{-5}	+76.4* [+68.2, +83.8]	−12.6* [−15.3, −10.0]	−76.3*	701 / 1,000	–	+16.6 → −0.1*
800 alone, 3 epochs, $3{\times}10^{-5}$	+76.5* [+68.3, +84.0]	−12.7* [−15.4, −10.1]	−76.8*	0 / 1,000	–	+16.6 → +0.3*

A weight-averaged repair (exploratory). Averaging the weights of the moderate-dose own-state model with the untouched student (0.6 and 0.4), with no decode-time intervention, raises delivered accuracy on MATH-500 from 79.3 to 81.8% (+2.5 $[+0.9, +4.1]$; four decoding seeds, one training seed) while capping falls from 15.7 to 11.0%, and also raises delivered accuracy on GPQA-

Diamond (18.1 → 20.0%, capping 28.3 → 12.1%) and LiveCodeBench (6.7 → 8.1%, capping 43.7 → 24.9%); on AIME 2025 accuracy is −2.5 [−11.7, +6.7] with capping 47.5 → 31.7%. On 44 problems outside both repair sets it still gains from a larger budget (8,192 → 16,384 tokens: delivered +4.0 [+1.1, +8.0]). These runs use a separate evaluation pass whose baseline differs slightly from the tables above.

A7 WHERE THE FORMAT GAP LANDS, AND WHAT EACH BRANCH LOOKS LIKE

Capping is the sum of a solved branch (budget reached with a correct last box) and a wrong branch (budget reached otherwise), and the wrong branch equals wrong×(1−stop-after-wrong); both hold by construction. The split is useful because it shows where each intervention acts. Table 20 splits every student's format gap. In every single-format student that breaks, most of the gap lies in the solved branch, from 64% (the 32B student) to 97% (the main student); in four of the five, deleting the system turn also lowers stop-after-wrong, and in none of the four students trained on two formats does it. The two branches also look different (main student, both decoding seeds). Under NO-SYSTEM, the 184 solved-then-capped generations write their first box at a median 11% of the budget, correct 97% of the time, and then loop verbatim ("Now produce final answer."), closing the thought in 2.7%; 151 of them stopped correctly under EXACT. Under EXACT, 85% of the 127 wrong-and-capped generations never write an answer at all, 92% are still-varying re-derivation rather than a loop, and 0.8% close the thought; wrong-and-capped generations under NO-SYSTEM look the same, so the behaviour follows the branch rather than the format (those problems are harder).

Table 20: **The format gap lies mostly in the solved branch, and two formats protect both branches.** MATH-500, NO-SYSTEM minus EXACT, points; † one decoding seed, otherwise two; * significant. Stop-after-wrong is $P(\text{natural stop} \mid \text{last box wrong})$ under each format.

student	format gap	Δ solved branch	Δ wrong branch	stop-after-wrong EXACT→NO-SYSTEM
Qwen3-8B-Base, maths (main)	+16.6*	+16.1*	+0.5	23.5→23.3
Qwen3-8B-Base, maths, 2nd training seed†	+29.0*	+26.2*	+2.8	27.4→14.4
Qwen3-8B-Base, science†	+55.4*	+36.6*	+18.8*	92.5→46.2
Qwen2.5-7B, maths	+40.0*	+37.7*	+2.3*	27.5→12.4
DS-R1-Distill-32B, maths†	+39.6*	+25.2*	+14.4*	33.3→4.9
Qwen3-8B-Base, maths, two formats	−1.1	−0.7	−0.4	25.3→24.6
Qwen2.5-7B, maths, two formats	−0.8	−0.8	+0.0	27.1→29.3
Qwen3-8B-Base, science, two formats	−2.3*	−0.3	−2.0*	89.4→92.7
Qwen3-8B-Base, maths, combined recipe	+0.9	−0.1	+1.0	25.0→26.8
Llama-3.1-8B base, maths	+0.0	+1.0	−1.0	36.2→38.9
Llama-3.1-8B-Instruct, maths†	+0.2	+1.8	−1.6	31.4→36.5
Mistral-7B-v0.3, maths	+0.7	−0.3	+1.0	30.2→30.3

Table 21: **Students of every family stop after a wrong answer less often than their bases, whether or not they break on (ii).** MATH-500 under each student's training format; stop-after-wrong in %, Δ its change from base to student, * significant. The Llama-3.1-8B and Mistral bases write no answer box in about 99% of rows, so their stop-after-wrong is their natural-stop rate; four bases were run on the first 100 problems (two decoding seeds), and each pair is matched on common problems.

			base				student				
initialization	student	problems	acc	no box	stop-after-wrong	cap	acc	stop-after-wrong	Δ	cap	breaks on (ii)
Qwen3-8B-Base	maths traces (main)	500	53.2	25.8	95.7	2.0	83.4	23.5	−72.2*	15.0	yes
Qwen3-8B-Base	science traces	500	53.2	25.8	95.7	2.0	49.0	92.5	−3.2	4.8	yes
Qwen3-8B-Base	code traces	500	53.2	25.8	95.7	2.0	51.6	52.5	−43.2*	29.0	no
Qwen2.5-7B	maths traces	100	31.5	39.5	99.3	0.5	78.5	30.2	−69.0*	16.5	yes
Qwen2.5-14B	maths traces (blind)	500	36.0	43.3	97.7	1.5	83.0	27.6	−70.0*	15.1	no
DS-R1-Distill-Qwen-32B	maths traces	500	85.7	7.3	65.0	5.4	92.2	33.3	−31.7*	10.0	yes
DS-R1-Distill-Llama-8B	maths traces (blind)	500	79.7	11.0	51.2	10.6	81.9	22.1	−29.1*	18.0	no
Llama-3.1-8B	maths traces	100	0.0	99.0	36.0	64.0	54.5	36.3	+0.3	33.0	no
Llama-3.1-8B-Instruct	maths traces	100	3.5	89.0	94.3	5.5	68.0	34.4	−59.9*	23.0	no
Mistral-7B-v0.3	maths traces	100	0.0	99.5	62.0	38.0	55.0	24.4	−37.6*	37.5	no

A8 CAUSE (I): THE ONE-VARIABLE TEST

Table 22: **Withholding terminal supervision destroys the end-of-turn token, and a standard stop set cannot see it.** Same base, rows, trainer, seed and step count; the arms differ only in whether each episode's terminal token stays in the supervised span. Under the stop set a standard harness uses the two checkpoints read alike on capping, accuracy and loops (well-formed stops rise); restricted to the end-of-turn token, every generation of the unsupervised arm runs to the budget. MATH-500, n=2,000 per arm.

	cap %	well-formed stop %	loop %	acc %	ended on <|im_end|> %	median gen. tokens
math500, trained format, harness stop set (500 items × 4 seeds)						
TERM	18.4	68.7	3.9	79.2	81.7	1,619
NOTERM	18.6	81.4	4.5	79.6	0.0	1,658
	+0.2	**+12.8**	+0.6	+0.3	**−81.7**	+12
difference	[−1.4, +2.0]	[+10.6, +14.8]	[−0.7, +1.8]	[−1.4, +2.0]	[−84.3, −79.0]	[−84, +110]
math500, trained format, stop only on <|im_end|> (500 items × 2 seeds)						
TERM	17.5	70.8	3.6	78.3	82.5	1,676
NOTERM	100.0	0.0	66.4	45.5	0.0	8,192
	+82.5	**−70.8**	**+62.8**	**−32.8**	**−82.5**	**+5,176**
difference	[+79.7, +85.3]	[−73.8, −67.8]	[+59.0, +66.4]	[−36.6, −29.2]	[−85.3, −79.7]	[+4,949, +5,405]
aime25, trained format, harness stop set (30 items × 8 seeds)						
TERM	51.7	43.3	18.3	22.9	48.3	30,000
NOTERM	51.2	48.8	22.1	25.4	0.0	30,000
	−0.4	+5.4	+3.8	+2.5	**−48.3**	+163
difference	[−7.9, +6.7]	[−2.1, +12.9]	[−3.3, +10.8]	[−2.9, +7.5]	[−61.7, −35.4]	[−1,552, +1,774]
math500, replication: TERM retrained in this session (500 items × 4 seeds)						
TERM	20.4	67.2	4.8	78.6	79.7	1,740
NOTERM	18.6	81.4	4.5	79.6	0.0	1,658
	−1.8	**+14.2**	−0.4	+1.0	**−79.7**	−85
difference	[−3.5, +0.1]	[+12.1, +16.4]	[−1.8, +0.9]	[−0.7, +2.8]	[−82.5, −76.8]	[−188, +17]

Table 23: **Deleting the system turn makes the end-of-turn-supervised arm stop emitting the end-of-turn token, as the end-of-turn-cut arm always does.** Under the harness stop set the cut arm, which already stops on the base's <|endoftext|>, shows about half the supervised arm's format gap (+28.0 against +52.2, a difference of −24.2, all of it in the solved branch); under the trained format the wrong branch differs by +0.9, not significant. This pair was trained separately from the main student and has a larger format gap; compare within the pair.

arm	format	cap	acc	stop after wrong	solved then capped	wrong and capped	ended on <|im_end|>
harness stop set (500 items × 4 seeds)							
TERM	trained	18.4	79.2	35.2	4.9	13.4	81.7
	no system	70.6	78.2	12.6	51.6	19.0	1.1
	gap	+52.2	−1.0	−22.5	+46.7	+5.5	−80.6
NOTERM	trained	18.6	79.6	29.4	4.2	14.4	0.0
	no system	46.6	76.5	14.2	26.4	20.2	0.0
	gap	+28.0	−3.1	−15.2	+22.2	+5.8	+0.0
stop only on <|im_end|> (500 items × 2 seeds)							
TERM	trained	17.5	78.3	37.8	4.0	13.5	82.5
	no system	97.9	56.5	0.9	54.8	43.1	2.1
	gap	+80.4	−21.8	−36.9	+50.8	+29.6	−80.4
NOTERM	trained	100.0	45.5	0.0	45.5	54.5	0.0
	no system	100.0	42.8	0.0	42.8	57.2	0.0
	gap	+0.0	−2.7	+0.0	−2.7	+2.7	+0.0

A9 SUPPORTING TABLES FOR CAUSE (II)

Table 24: **Taking both closing coordinates from the perturbed format reproduces the failure.** The reverse arm of §4, run on both closing tokens rather than </think> alone. Under the training format with both coordinates replaced, capping rises $18.2 \rightarrow 49.0\%$ with the phenotype of the real perturbation (solved-then-capped $4.4 \rightarrow 32.0$, loops $1.2 \rightarrow 34.4$). The overshoot against the 13.4 points the perturbation itself costs in this cohort is what shows the reasoning conditionals are not inert. MATH-500, n=500. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
EXACT (plain sampling)	500	18.2	82.2	4.4	1.2	81.8	0.0
NO-SYSTEM (plain sampling)	500	31.6 (+13.4*)	79.6 (−2.6)	14.6 (+10.2*)	14.2 (+13.0*)	55.6 (−26.2*)	12.8 (+12.8*)
EXACT + </think> and end-of-turn from NO-SYSTEM (reverse, both)	500	49.0 (+30.8*)	78.6 (−3.6*)	32.0 (+27.6*)	34.4 (+33.2*)	42.2 (−39.6*)	8.8 (+8.8*)
NO-SYSTEM + </think> and end-of-turn from EXACT (rescue)	500	17.8 (−0.4)	81.0 (−1.2)	3.4 (−1.0)	2.0 (+0.8)	82.2 (+0.4)	0.0

Table 25: **Two formats from step 0 prevent the dependence; adding an unseen turn does not create it.** Two students trained for this paper with the main student's exact recipe (Qwen3-8B-Base, the same 30,000 traces, seed, optimizer and 855 steps), differing only in how the training rows are formatted. **ARM 1** carries the system turn on a seeded half of the rows and none on the other half; **ARM 2** carries it on none. Both predictions were pre-registered before either job was created and the file is in the supplementary material: ARM 1 was predicted not to meet the endpoint and does not; ARM 2 was predicted to meet it in the mirror direction and *does not*, which we report as the falsification it is. † marks formats that appear in no training arm. MATH-500, budget 8,192; the main student's row is the paper's published cell for reference. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding format	n	cap	acc	solved-then-capped	loop	well-formed	malformed
main student: trained EXACT *only (§2)*							
EXACT (own training format)	1000	15.0	83.4	2.3	1.3	85.0	0.0
NO-SYSTEM (the endpoint perturbation)	1000	31.6 (+16.6 [+13.2, +20.0]*)	82.8 (-0.6)	18.4 (+16.1*)	16.3 (+15.0*)	54.4 (-30.6*)	14.0 (+14.0*)
***ARM 1**: trained* EXACT + NO-SYSTEM, *50/50 from step 0*							
EXACT (own training format)	1000	16.4	82.6	3.4	1.2	83.6	0.0
NO-SYSTEM (the endpoint perturbation)	1000	15.3 (-1.1 [−3.0, +0.8])	83.3 (+0.7)	2.7 (-0.7)	0.4 (-0.8)	84.7 (+1.1)	0.0 (+0.0)
INSTR	500	15.0 (-1.8)	83.2	–	–	–	–
HISTORICAL	500	13.4 (-3.4*)	82.6	–	–	–	–
ALT-SYSTEM†	500	16.0 (-0.8)	82.0	–	–	–	–
EMPTY-SYSTEM†	500	17.2 (+0.4)	80.8	–	–	–	–
***ARM 2**: trained* NO-SYSTEM *only (mirror direction)*							
NO-SYSTEM (own training format)	1000	15.4	82.0	2.5	1.1	84.6	0.0
EXACT (the endpoint perturbation)	1000	16.8 (+1.4 [−0.5, +3.3])	81.7 (-0.3)	2.9 (+0.4)	1.4 (+0.3)	83.2 (-1.4)	0.0 (+0.0)
INSTR	500	14.4 (-1.2)	85.4	–	–	–	–
HISTORICAL	500	15.4 (-0.2)	82.0	–	–	–	–
ALT-SYSTEM†	500	16.2 (+0.6)	82.6	–	–	–	–
EMPTY-SYSTEM†	500	15.8 (+0.2)	83.2	–	–	–	–
The endpoint at the intermediate checkpoints (epoch 1 and 2 of the same runs; n=1000 each)							
ARM 1, epoch 1: cap EXACT → NO-SYSTEM	1000	17.5 → 14.6 (-2.9 [−5.1, −0.7]*)	81.2	–	–	–	–
ARM 1, epoch 2: cap EXACT → NO-SYSTEM	1000	16.6 → 16.3 (-0.3 [−2.4, +1.8])	82.6	–	–	–	–
ARM 2, epoch 1: cap NO-SYSTEM → EXACT	1000	15.3 → 14.6 (-0.7 [−2.8, +1.4])	82.1	–	–	–	–
ARM 2, epoch 2: cap NO-SYSTEM → EXACT	1000	16.2 → 16.5 (+0.3 [−1.8, +2.4])	81.8	–	–	–	–

Table 26: **The two-format repair holds off MATH-500.** Format gap (NO-SYSTEM−EXACT capping, points) of the single-format reference and of the student trained on two formats (or the combined recipe of §6), same benchmark, budget and decoding seeds; * significant. GPQA accuracy changes are omitted: the letter judge used for them undercounts correct multiple-choice answers by about half, whereas capping does not depend on the judge.

student	benchmark	n	single-format gap	repaired gap	repaired Δacc, NO-SYSTEM−EXACT
Qwen3-8B-Base, two formats	GPQA-Diamond	792	+28.0*	+0.4 [−3.4, +4.0]	–
Qwen3-8B-Base, two formats	AIME 2025	360	+17.5*	−2.5 [−8.3, +2.8]	+0.3
Qwen3-8B-Base, combined recipe	GPQA-Diamond	792	+28.0*	+1.9 [−1.4, +5.2]	–
Qwen3-8B-Base, combined recipe	AIME 2025	360	+17.5*	+1.4 [−4.2, +6.9]	−1.1
Qwen2.5-7B, two formats	GPQA-Diamond	792	+13.9*	+0.9 [−2.5, +4.5]	–
Qwen2.5-7B, two formats	AIME 2025	240	+7.1	−1.7 [−8.3, +5.0]	−1.7
Qwen3-8B-Base science, two formats	GPQA-Diamond	792	+45.3*	−1.0 [−5.8, +4.0]	–

Table 27: **Shift over margin at the closing tokens separates every case that breaks from every case that holds, and quantifies the deletion/addition asymmetry.** Teacher-forced at the closing positions; *own stops* are the student's own natural stops under its training format, *held-out traces* the teacher traces of Appendix A14.1. *margin* is the closing token's logit minus its strongest competitor's; *shift* is the margin lost under the perturbation and *shift/margin* its fraction. The two cells that break lose 0.90 and 1.64 of their margin; the eight that hold lose at most 0.43. On the pretrained bases the closing token is not yet the argmax, so the ratio is undefined there and the shift is read as a displacement (Appendix A4). The last four rows are the pre-registered arms of Table 25: the two-format student's margin does not move under either format, and the mirror student (‡: columns printed own → perturbed, i.e. NO-SYSTEM → EXACT) loses 2.4 of 6.3 logits when a turn it never saw is added—about half of what the main student loses when its turn is deleted—and keeps stopping.

checkpoint	closing token	n	p(closing token) EXACT → NO-SYSTEM	logit margin to best competitor, EXACT → NO-SYSTEM	logit shift	shift / margin	% argmax EXACT → NO-SYSTEM	
Qwen3-8B CE student, own EXACT stops	</think>	150	$0.995 \to 0.113$	$+6.38 \to +0.62$	5.75	0.90	$100 \to 70$	**breaks**
same student, end-of-turn	end-of-turn	150	$0.989 \to 0.828$	$+5.50 \to +3.38$	2.12	0.39	$100 \to 98$	holds
Qwen2.5-7B CE student, own EXACT stops	</think>	150	$1 \to 1$	$+12.88 \to +12.12$	0.75	0.06	$100 \to 100$	holds
same student, end-of-turn	end-of-turn	150	$0.997 \to 8.3 \times 10^{-4}$	$+8.81 \to -5.62$	14.44	1.64	$100 \to 6$	**breaks**
32B CE student, held-out traces	</think>	182	$1 \to 1$	$+10.19 \to +10.12$	0.06	0.01	$100 \to 100$	holds
same student, end-of-turn	end-of-turn	182	$0.983 \to 0.981$	$+4.75 \to +4.50$	0.25	0.05	$97 \to 97$	holds
Qwen3-8B-Base, before SFT	</think>	182	$2.3 \times 10^{-7} \to 1.1 \times 10^{-8}$	$-12.84 \to -16.50$	3.66	–	$0 \to 0$	–
same base, end-of-turn	end-of-turn	182	$5.1 \times 10^{-9} \to 2.1 \times 10^{-9}$	$-18.75 \to -19.62$	0.88	–	$0 \to 0$	–
Qwen2.5-7B, before SFT	</think>	182	$0.013 \to 0.007$	$-2.00 \to -3.06$	1.06	–	$6 \to 1$	–
same base, end-of-turn	end-of-turn	182	$2.5 \times 10^{-10} \to 1.3 \times 10^{-11}$	$-21.97 \to -24.88$	2.91	–	$0 \to 0$	–
ARM 1 (both formats), own EXACT stops	</think>	150	$0.994 \to 0.994$	$+6.31 \to +6.25$	0.06	0.01	$100 \to 100$	holds
same student, end-of-turn	end-of-turn	150	$0.988 \to 0.988$	$+5.38 \to +5.25$	0.12	0.02	$100 \to 100$	holds
ARM 2 (NO-SYSTEM only), own stops‡	</think>	150	$0.994 \to 0.951$	$+6.31 \to +3.94$	2.38	0.38	$100 \to 100$	holds
same student, end-of-turn	end-of-turn	150	$0.988 \to 0.927$	$+5.38 \to +3.06$	2.31	0.43	$100 \to 100$	holds

Under the training format — stops after $6{,}890$ characters:

`...Thus` $\angle AHB = 180^\circ - 74^\circ = 106^\circ$`.` `\boxed{106^\circ} [end of turn]`

Same weights, same problem, same seed, system turn deleted — reaches the correct answer *sooner*, at character $1{,}405$, then does not stop:

`\boxed{106^\circ}.` **`But let's reason step by step.`** `We can also derive using geometry: Since AD is altitude ...`

`...` [18,264 further characters of re-derivation, decaying into] `...D_x = 0.793565057788928. Then D_y = m_AD * D_x = 0.`

budget exhausted at $19{,}669$ *characters.*

Figure 7: **One problem, both formats.** The perturbation removes eleven tokens that say nothing about the task. The model does not become worse at geometry—it reaches the same correct answer, and reaches it earlier—but having reached it, it re-opens the derivation and never closes. This is the modal failure: across MATH-500, 103 of the 165 budget-capped generations under the perturbed format already contain a correct boxed answer.

Table 28: **The stop decision breaks where the teacher stops, and nowhere else.** The continuation history is held fixed and only the prompt prefix differs, so any change is attributable to the prefix alone. Qwen3-8B-Base student, 250 teacher trajectories from the training traces (an on-manifold control: the teacher's own stopping point is a state the student was supervised on). The two controls are what make the first rows interpretable: mid-thought the student is *equally* reluctant to close under both formats, so it has not become trigger-happy; after a generation has already run to the budget neither format can close it, so the damage is not repairable late.

where we probe (history fixed, only the prefix differs)	quantity	EXACT	NO-SYSTEM
teacher's terminal state, answer written	P(end of turn), median	0.975	$\mathbf{0.739}$
same state, 16-token greedy continuation	ends the turn	98.0%	$\mathbf{90.0}\%$
same state, 8 sampled continuations	ends the turn	97.0%	$\mathbf{88.4}\%$
control: teacher's first correct answer, mid-thought	P(close thought)	2.3%	2.3%
control: end of a budget-capped generation	P(stop or close)	$<10^{-6}$	$<10^{-6}$

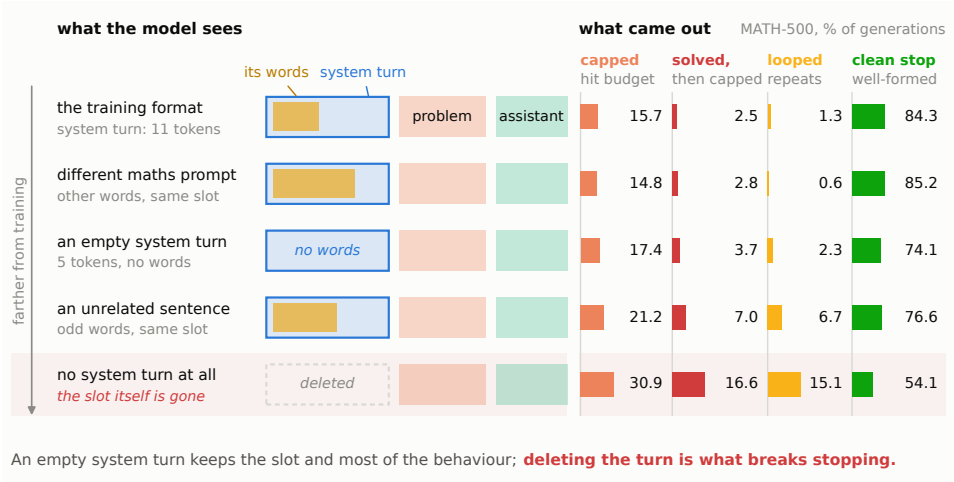

Figure 8: **The presence of the system turn carries most of the effect; its words carry the rest.** Rows: prompt formats by distance from the training format, as the chat template emits them—the system turn is a blue *slot* (its framing tokens) holding an amber bar for its words, then problem and assistant turns; widths schematic, a deleted block ghosted. Right: MATH-500 outcomes, each column on its own scale. *capped* hit the 8,192-token budget, *solved-then-capped* capped with a correct last boxed answer, *looped* repeated, *clean stop* stopped naturally with the thought closed. n=1000, two decoding seeds (n=500 for the different maths prompt); all cells, with AIME, in Table 66.

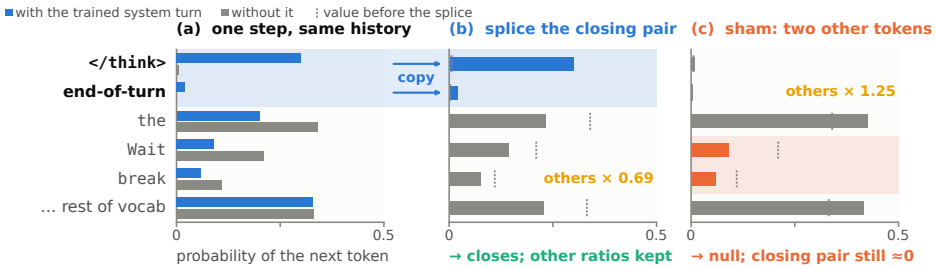

Figure 9: **What the splice does, at one decoding step.** **(a)** The next-token distribution with the trained system turn (blue) and without it (grey), given the same generated history, at a step where the training format would close. **(b)** What we sample: the two closing coordinates take their training-format values (*copy*), and every other coordinate keeps its NO-SYSTEM value times one common factor, so every ratio among them is unchanged—the arm adds closure mass and reorders nothing else. It is not "sample under the training format": away from the spliced pair the ordering stays the NO-SYSTEM one. **(c)** The sham performs the identical operation on two other tokens; it moves *more* mass than (b) and is null, because the closing pair is left at its perturbed value. Bar heights are schematic; the substitution and both rescaling factors are computed from Eq. 1.

Table 29: **Where the fragility lives.** TinyR1-32B-Preview and the checkpoint it was fine-tuned from, crossed over the one feature separating their templates: whether the assistant turn prefills a thought opener. Along each diagonal the prompts are byte-identical (SHA-256 verified before any generation), so those pairs differ only in weights. *cap, own* is capping under the row's own prompt; Δ is the paired change when a generic system turn is added. Only the child's native cell breaks. Its no-prefill cell starts from a much worse baseline (26.4% capping), which makes a further rise harder to detect, though at that rate it is not at ceiling. MATH-500, n=500 per cell; all four possible readings were registered before these rows existed.

weights	prompt	cap, own	Δcap	Δacc	verdict
parent	no prefill *(its own)*	2.0	-0.2 $[-1.8, +1.4]$	-4.2	flat
parent	with prefill	2.0	-0.6 $[-2.0, +0.8]$	-1.0	flat
child	no prefill	**26.4**	$+2.8$ $[-2.0, +7.8]$	$+1.0$	already failing
child	with prefill *(its own)*	7.2	$+$**11.0** $[+7.4, +14.6]$	$+0.0$	**breaks**

Table 30: **Training choices, with the number of independent training runs behind each.** MATH-500, n=500 per cell at two decoding seeds; all optimizer pairs measured in one runtime, which also removes a seed/hardware confound. *seed spread* is the min–max of the perturbed-format gap across training seeds of the same recipe; the reference band from four seeds of one recipe is 16.5 points, so a claim is only separable from training noise if it clears that.

training choice	setting	steps	**runs**	cap, own → perturbed	loops, own → perturbed	gap [min,max]
cohort A: CE objective, standard AdamW, 30k problems unless stated; perturbed = NO-SYSTEM						
epochs	1× 30k	285	1	19.2 → 67.7	3.9 → 48.0	48.5 single run
epochs	3× 30k	855	4	14.4–16.7 → 31.1–47.1	0.8–1.8 → 15.1–27.6	**24.6** [14.4, 30.9]
problem diversity	1× 100k distinct	937	2	14.8–15.8 → 49.3–61.2	1.5 → 31.2–43.0	**40.0** [33.5, 46.4]
cohort B: optimizer, pooled over three objectives × two training seeds; perturbed = NO-SYSTEM						
optimizer	standard AdamW $(0.9, 0.95)$	855	6	14.4–18.5 → 31.1–66.9	0.7–5.1 → 15.1–46.6	**31.7** [14.4, 51.9]
optimizer	β_1=β_2=0.95	855	6	15.4–17.4 → 15.0–25.6	0.8–3.9 → 0.9–10.3	**3.0** $[-1.2, 10.2]$
cohort C: objective, perturbed = HISTORICAL						
objective	cross-entropy	855	2	14.4–16.7 → 25.8–35.1	0.8–1.8 → 13.1–21.9	**14.9** [9.1, 20.7]
objective	confidence-weighted CE	855	2	18.4–18.5 → 48.2–49.2	4.3–5.1 → 36.0–37.3	**30.2** [29.7, 30.8]

A10 How SFT stacks guard the terminal token

We read the packing and truncation paths of widely used SFT stacks and asked one question of each: can a sequence longer than the configured maximum lose its terminal token from the *supervised* span, and is that loss silent?

Three projects defend the boundary, each in its own way. ms-swift, by default, drops a row longer than the maximum instead of cutting it, and a separate pass supervises the separator after every assistant turn, which for ChatML is the end-of-turn token itself; its explicit left or right truncation cuts the end-of-turn token like any other token and nothing re-attaches it. Axolotl exposes a flag that puts the end-of-sequence token under loss. Megatron-LM forces an end-of-document token when a sample is cut. No two of these are the same mechanism, and none verifies the data it emits; the defence is a code path, not a check.

Two gaps follow. Megatron-LM's guard is wired for one of its two supported prompt formats, so the failure survives in the other. And the pattern that newer stacks prefer—letting the chat template mark the supervised region with a `{% generation %}` block—has no defence at all when the template does not mark one; of the templates we hold, none does. ms-swift's default is safe for the end-of-turn token but silently shrinks the dataset, by 85% of our rows at a 4k budget: the loss is visible only as a difference in row counts.

How often the end-of-turn token is cut. Table 31 runs both stacks' own encoding on our $30{,}000$ episodes and on the public s1K-1.1 set ($1{,}000$ episodes, same format and tokenizer). TRL keeps the first L tokens of every row, without a warning, packed or not; at its default L=$1{,}024$ only 0.19% of our episodes and 0.1% of s1K-1.1 keep their end-of-turn token under loss, and even at 16k, 14.4% and 10.2% lose it. The cut lands inside the thought: `</think>` is the last supervised token in at most 0.01% of any cell, so a student trained this way sees long thoughts that neither close nor end. The prevalence is fixed by one comparison, row length against L, which is why the check belongs in the data path and not in the stack.

Table 31: **Under TRL's defaults the terminal token leaves the supervised span exactly as often as rows exceed the budget; ms-swift keeps it by dropping those rows.** Share of supervised episodes whose last supervised token is `<|im_end|>`; Qwen3-8B-Base tokenizer, the paper's format; packing on and off agree in every cell. ms-swift 3.12 (`truncation_strategy=delete`), TRL 0.24 (`max_length` truncation).

		ms-swift default: drop the row			TRL default: keep the first L tokens (no row dropped)			
data	L	episodes	terminal %	dropped %	episodes	terminal %	ends on `</think>` %	`</think>` supervised %
our set (30k)	4k	4,473	100	85.1	30,000	14.9	< 0.1	21.1
	8k	13,747	100	54.2	30,000	45.8	< 0.1	51.6
	16k	25,685	100	14.4	30,000	85.6	0	89.0
	32k	29,991	100	< 0.1	30,000	> 99.9	0	> 99.9
s1K-1.1 (1k)	4k	96	100	90.4	1,000	9.6	0	13.5
	8k	387	100	61.3	1,000	38.7	0	43.3
	16k	898	100	10.2	1,000	89.8	0	93.0
	32k	1,000	100	0	1,000	100	0	100

The common failure is not that maintainers are unaware of the boundary. It is that every defence is written once, per stack, and nothing downstream confirms it worked on the data actually produced. That is the gap our check fills.

The agent-trace set of Section 3 lost its end-of-turn token through none of these packing paths: it was neither packed (one episode per row) nor truncated. Our formatter emitted every end-of-turn marker with loss disabled and the converter copied that mask, so `<turn|>` occurred $651{,}849$ times in the inputs and was supervised 0 times; a rebuild that supervises the marker after each assistant span, with every token id unchanged, restores it in all $287{,}735$ turns. Loaded at $131{,}072$ tokens, the rebuilt set also truncated 92 rows inside the supervised span, losing $2{,}210$ end-of-turn tokens. It is a different corpus, model (Gemma-4) and stack (Megatron-Bridge) from the mathematics students, whose $30{,}000$ episodes each supervise their end-of-turn token once.

A11 TERMINATOR: THE DECODE GUARD IN DETAIL

The decode guard inspects every still-running generation once, at a fixed token count T: it delivers the boxed answer the thought already holds, or forces an answer where it holds none, and ends the turn. It removes capping, raises computed and delivered accuracy for every student in both formats, leaves the base unchanged and shortens generations (Table 32; AIME 2025, GPQA-Diamond, the released checkpoints and s1's budget forcing: Table 4); the end-of-turn-cut student is the one case it cannot repair. Rule, choice of T, the check of the implementation and all five predictions were pre-registered before any job ran; every prediction holds.

The rule. Decoding is the harness's (temperature 0.6, top-p 0.95, per-request seed, stop set, budget B=8,192 on MATH-500); the guard adds only the following, where `EOT` is the format's end-of-turn token and `x` the last boxed answer:

```
at check point T (then T+512, T+1024, ... < B), for each generation that has not stopped:
  p = its first T generated tokens, decoded
  if the thought in p is still open:
    if p contains \boxed{x}:                                        [rule (a), no model call]
      append "\n</think>\n\n**Answer:** \boxed{x}" + EOT
    else:                                                           [rule (b)]
      append "\n\nTime is up; I must commit to my best answer now."
           + "\n</think>\n\n**Answer:** \boxed{"
      sample <= 64 tokens with the same sampler, seed and stop set
      if none of them is a stop token: append EOT                   [forced close]
  else (thought closed, turn still running; rare):
    box after the closer:      append EOT                           [a_vis]
    box only in the thought:   append "\n\n**Answer:** \boxed{x}" + EOT   [a_th]
    no box:                    the rule-(b) string without </think>  [b_closed]
```

Both rules end every generation at T, so the later check points only ever meet a commit that sampled 64 tokens without stopping, and the forced close has already ended it; one check suffices, and a generation uses at most T tokens plus the commit string, 64 sampled tokens and `EOT`, about $T{+}100 \leq B$. Nothing is written on the student's behalf: the delivered answer is the last box of its own thought (a) or what it samples after the answer prompt (b), and rule (b) leaves a boxed answer after the closer in 94.6% of the main student's own-format commits.

Table 32: **The guard raises computed and delivered accuracy in every student cell, leaves the base unchanged and shortens generations.** MATH-500, budget 8,192, T=6,144; untouched → guard (paired change); capping is zero under the guard by construction; mean tokens include the appended strings. Science and 32B rows pool decoding seeds 1 and 2; replicate decode: the same checkpoint decoded on the second installation (Appendix A12.3); ‡ decoded with <|im_end|> as the only stop token; * significant.

student	format	n	untouched cap	accuracy	delivered	mean tokens
main student	EXACT	1000	15.0	83.4 → 85.7 (+2.3*)	80.7 → 85.3 (+4.6*)	2820 → 2489
main student	NO-SYSTEM	1000	31.6	82.8 → 85.7 (+2.9*)	51.5 → 73.2 (+21.7*)	3734 → 3082
replicate decode	EXACT	1000	15.7	82.3 → 85.2 (+2.9*)	79.6 → 85.0 (+5.4*)	2834 → 2493
replicate decode	NO-SYSTEM	1000	30.9	82.2 → 84.3 (+2.1*)	51.6 → 70.5 (+18.9*)	3699 → 3056
science student	EXACT	1000	5.8	46.2 → 47.0 (+0.8*)	43.6 → 45.0 (+1.4*)	1539 → 1419
science student	NO-SYSTEM	1000	60.4	56.4 → 65.0 (+8.6*)	13.4 → 58.1 (+44.7*)	5353 → 4133
32B student	EXACT	1000	10.5	90.9 → 93.2 (+2.3*)	86.7 → 93.0 (+6.3*)	2603 → 2356
32B student	NO-SYSTEM	1000	48.9	78.9 → 92.0 (+13.1*)	50.0 → 92.0 (+42.0*)	5350 → 4325
Qwen3-8B-Base	EXACT	500	2.0	53.2 → 53.6 (+0.4)	–	731 → 691
terminator cut‡	EXACT	1000	100.0	45.5 → 49.5 (+4.0*)	0.0 → 49.5 (+49.5*)	8192 → 6157
supervised twin‡	EXACT	1000	17.5	78.3 → 81.0 (+2.7*)	64.1 → 71.0 (+6.9*)	3016 → 2636

Replay, checked against native decoding. The first T tokens of each generation are replayed from the stored untouched generation (same checkpoint, format, sampler, seed and stop set, token ids kept per row). A chunked decode whose first chunk is T tokens produces exactly such a prefix, so guard and untouched decoding are compared on the same sample paths and rows that ended before T are identical by construction. As an implementation check we also ran the guard natively, generating the first chunk afresh, on the main student for decoding seed 1, both formats and every T (Table 33). vLLM's batch nondeterminism reproduces the stored prefix on only 11–13 of 500 items, so agreement is distributional: native minus replay is between -2.8 and $+2.0$ points in delivered

accuracy and between -1.4 and $+1.6$ in computed accuracy, every 95% CI contains zero, and the two rules fire at similar rates.

Choice of T. T was chosen on one cell, the main student under EXACT with decoding seed 1 (3102801): the $T \in \{2{,}048, 4{,}096, 6{,}144\}$ with the highest delivered accuracy, the larger one if two are within 0.5 points. Delivered accuracy is 81.0, 84.6 and 84.8%, so T=6,144. Seed 2 (3102802) is the held-out read: 81.2% untouched, 80.8, 83.6 and 85.8% under the guard, the same order (Table 34, which gives every cell at all three T). Early checks cost computed accuracy under EXACT, 83.4 → 81.0% on the main student and 92.2 → 84.6% on the 32B student (seed 1) at T=2,048, because they cut generations that would still have solved; without the system turn delivered accuracy is already near its plateau there (71.4, 73.0, 73.2% on the main student).

Table 33: **Native decoding agrees with the replay within noise.** Main student, decoding seed 1, n=500 per row: the guard with a freshly generated first chunk (native) against the guard on the stored untouched prefix (replay). Differences are native minus replay with 95% CIs; *fires*: generations handled by rule (a) or (b), native / replay; *same prefix*: native first chunks identical to the stored tokens.

		delivered %		accuracy %		(a)	(b)	same
format	T	native / replay	difference	native / replay	difference	fires	fires	prefix
EXACT	2,048	82.2 / 81.0	+1.2 [−1.6, +4.0]	82.2 / 81.0	+1.2 [−1.6, +4.0]	71 / 75	127 / 130	12
EXACT	4,096	85.6 / 84.6	+1.0 [−1.8, +4.0]	85.6 / 85.0	+0.6 [−2.2, +3.4]	37 / 38	83 / 82	12
EXACT	6,144	86.8 / 84.8	+2.0 [−0.8, +4.8]	86.8 / 85.2	+1.6 [−1.2, +4.4]	33 / 26	66 / 67	12
NO-SYSTEM	2,048	68.4 / 71.2	−2.8 [−7.2, +1.4]	79.2 / 80.6	−1.4 [−4.2, +1.4]	110 / 112	140 / 131	13
NO-SYSTEM	4,096	71.4 / 73.2	−1.8 [−6.6, +2.8]	84.6 / 84.4	+0.2 [−2.6, +3.0]	106 / 115	89 / 81	12
NO-SYSTEM	6,144	72.0 / 74.0	−2.0 [−6.8, +2.6]	86.0 / 85.4	+0.6 [−2.2, +3.4]	103 / 111	69 / 67	11

Table 34: T=6,144**, chosen on one seed of one cell, is also best on the held-out seed; early checks cut generations that would still have solved.** Delivered and computed accuracy (%) untouched and under the guard at each first check point T; seeds 1, 2: decoding seeds 3102801, 3102802; science and 32B cells are listed per seed. Base delivered accuracy is near zero by definition (it never closes a thought). ‡ as in Table 32.

			delivered %				accuracy %			
student	format	seeds	untouched	2,048	4,096	6,144	untouched	2,048	4,096	6,144
main student	EXACT	1,2	80.7	80.9	84.1	85.3	83.4	81.0	84.4	85.7
seed 1 (selects T)	EXACT	1	80.2	81.0	84.6	84.8	83.0	81.0	85.0	85.2
seed 2 (held out)	EXACT	2	81.2	80.8	83.6	85.8	83.8	81.0	83.8	86.2
main student	NO-SYSTEM	1,2	51.5	71.4	73.0	73.2	82.8	81.1	84.9	85.7
replicate decode	EXACT	1,2	79.6	80.2	84.2	85.0	82.3	80.2	84.4	85.2
replicate decode	NO-SYSTEM	1,2	51.6	70.1	70.3	70.5	82.2	80.5	83.6	84.3
science student	EXACT	1	48.0	48.0	48.6	49.4	49.0	48.0	48.6	49.4
science student	EXACT	2	39.2	41.0	41.2	40.6	43.4	44.6	45.0	44.6
science student	NO-SYSTEM	1	15.0	59.2	58.6	58.8	58.0	64.0	64.0	64.2
science student	NO-SYSTEM	2	11.8	56.8	57.4	57.4	54.8	64.2	65.6	65.8
32B student	EXACT	1	87.2	84.6	91.0	93.6	92.2	84.6	91.2	93.8
32B student	EXACT	2	86.2	85.6	90.0	92.4	89.6	85.6	90.2	92.6
32B student	NO-SYSTEM	1	49.4	85.2	90.0	92.0	79.4	85.2	90.0	92.0
32B student	NO-SYSTEM	2	50.6	84.0	89.4	92.0	78.4	84.0	89.4	92.0
replicate decode	EXACT	1	86.0	85.4	89.6	91.2	89.8	85.4	89.8	91.4
replicate decode	NO-SYSTEM	1	48.8	85.0	91.0	93.0	79.4	85.0	91.0	93.4
Qwen3-8B-Base	EXACT	1	–	–	–	–	53.2	53.8	53.8	53.6
terminator cut‡	EXACT	1,2	0.0	52.8	51.1	49.5	45.5	52.8	51.1	49.5
supervised twin‡	EXACT	1,2	64.1	68.3	70.2	71.0	78.3	76.8	79.4	81.0

Predictions and verdicts. Paired by (problem, seed), item-clustered bootstrap, at T=6,144:

- **P1**, capping goes to zero in every cell: holds, by construction.
- **P2**, cause (ii): the main student's NO-SYSTEM delivered accuracy rises (51.5 → 73.2%, +21.7, 95% CI [18.7, 24.6]) and its computed-minus-delivered gap shrinks by at least half of its capped part. The gap falls from 31.3 to 12.5 points; the capped part, 18.4, goes to

zero; the residual is almost entirely (12.4) stops that never closed the thought, which the rule never touches, as stated in advance. Holds.

- **P3**, cause (iii): under EXACT computed accuracy does not fall (lower CI bound above -2) while wrong-and-capped goes to zero, and delivered accuracy does not fall. Accuracy $+2.3$ ([1.0, 3.6]), wrong-and-capped $12.7 \to 0$, delivered $80.7 \to 85.3\%$. Holds.
- **P4**, harmlessness: the base's accuracy moves within ± 1 point with a CI reaching zero; $+0.4$ ([0.0, 1.0]), the guard firing on 10 of 500 generations. Holds; delivered accuracy is near zero for the base by definition, since it never closes a thought.
- **P5**, cause (i): the end-of-turn-cut student, decoded with the end-of-turn token as its only stop, goes from zero delivered accuracy (every generation capped) to within 10 points of its accuracy; it delivers 49.5%, equal to its accuracy under the guard. Holds as written; the paragraph on cause (i) below shows why this is still a failure.

Where each rule fires. On the main student's own format rule (a) fires on 5.8% of generations and delivers a correct answer on 63.8% of them, where the untouched run delivered 43.1%; rule (b) fires on 13.0%, and its forced answers are right 30.8% of the time against 11.5% computed (4.6% delivered) if left to run (Table 35). Of the 127 own-format wrong-and-capped generations, 24.4% end with a delivered correct answer; rule (b) handles 114 of them and rule (a) 13. Rule (a) has a small price: on its rows the untouched runs are correct 67.2% of the time by the last box, since a few generations holding a wrong box at T would still have corrected it; over the whole cell the guard gains 49 delivered correct answers and loses 3 (31 and 8 computed). Without the system turn rule (a) carries the effect: it fires on 21.5% and lifts delivered accuracy on those rows from 4.2 to 86.0%, because they are mostly the 18.4% solved-then-capped generations, 98.4% of which the guard delivers. The student usually writes on after its forced answer: 78% of its own-format commits end by the forced close.

Table 35: **Rule (a) delivers answers the student already holds; rule (b) forces answers that are right far more often than the same generations left to run.** Same cells as Table 32. *fires*: % of generations; (a) *delivered correct*: delivered accuracy of those generations untouched → under the guard; (b) *correct*: computed accuracy left to run → delivered accuracy of the forced answer; *forced close*: % of (b) commits that sampled 64 tokens without stopping; *wrong-and-capped*: the untouched wrong-branch generations and the % the guard turns into a delivered correct answer.

		(a) holds a box		(b) holds none			wrong-and-capped	
student	format	fires	delivered correct	fires	correct	forced close	n	delivered
main student	EXACT	5.8	43.1 → 63.8	13.0	11.5 → 30.8	78	127	24.4
main student	NO-SYSTEM	21.5	4.2 → 86.0	12.5	9.6 → 33.6	79	132	24.2
replicate decode	EXACT	4.9	32.7 → 65.3	13.9	9.4 → 30.9	78	132	27.3
replicate decode	NO-SYSTEM	19.7	7.1 → 85.3	13.7	10.9 → 28.5	84	143	20.3
science student	EXACT	1.3	15.4 → 53.8	5.4	3.7 → 20.4	94	52	19.2
science student	NO-SYSTEM	44.0	0.0 → 82.5	16.9	0.0 → 49.7	100	244	35.7
32B student	EXACT	6.5	50.8 → 84.6	8.4	28.6 → 56.0	88	65	36.9
32B student	NO-SYSTEM	28.1	9.3 → 99.3	24.4	18.0 → 72.1	22	201	67.7
Qwen3-8B-Base	EXACT	0.4	0.0 → 0.0	1.6	0.0 → 25.0	25	10	20.0
terminator cut‡	EXACT	86.4	0.0 → 52.3	13.6	10.3 → 31.6	100	545	10.1
supervised twin‡	EXACT	8.0	16.2 → 63.8	12.5	6.4 → 28.0	99	135	22.2

Token savings. Mean generated tokens, appended strings included, fall in every cell (Table 32): by 330 (12%) on the main student under EXACT and 653 (17%) under NO-SYSTEM, by 1,025 and 1,220 on the 32B and science students without the system turn, and by 40 on the base. The saving is the budget that generations still running at T would have spent after it.

Cause (i) stays upstream of the guard. The end-of-turn-cut student, decoded with <|im_end|> as its only stop, caps on every generation; the guard ends all of them and delivers 49.5%, while its supervised twin under the same stop set and guard delivers 71.0% (81.0% computed). By T the cut student has already emitted <|endoftext|>, which this stop set ignores, and generated past it: 846 of its 1,000 generations had closed the thought and written the answer before T (rule a_vis). The guard only truncates text the student should never have written, and an earlier check does not close the gap (52.8% delivered at T=2,048). The data guard has to come first.

The science student's first token. Under EXACT the science student's first generated token is set by the decoding seed on every item: with seed 1 all 500 generations open with `</think>`, closing the thought before it starts, and with seed 2 all 500 open with `<think>`. The two seeds are therefore two decoding modes rather than two draws, and its cells pool both. With seed 1 the guard can fire only through the closed-thought rules (all 10 rule-(a) firings are a_vis, all 22 rule-(b) firings b_closed). Under EXACT this student caps on only 5.8% and the guard adds 1.4 delivered points; without the system turn it lifts delivered accuracy by 43.8 and 45.6 points on the two seeds alike. Its NO-SYSTEM computed accuracy is the higher one (56.4 against 46.2% untouched), so under the guard its format gap turns negative (58.1 against 45.0% delivered).

The commit-trigger study behind rule (b). A companion study asked whether a trigger applied only to *stuck* generations, those that have neither stopped nor written `\boxed{` in their first T tokens, removes the main student's wrong branch without costing solved problems. It uses the same untouched cells; a stuck generation is cut at its last newline at or before T (back-off at most 400 tokens), a string is appended, and it continues with the same sampler and stop set within the budget. The arms, at $T \in \{4{,}096, 6{,}144, 7{,}168\}$: `resume`, a placebo that cuts and resamples; `ws`, a logit bias of -100 on the first tokens of *Wait*, *wait* and *Alternatively*, with and without a leading space, for the rest of the generation; `c1`, a commit inside the open thought (`\n\nI need to commit to an answer now. **Final answer:** \boxed{`); `c2`, a close and answer (`\n</think>\n\n**Answer:** \boxed{`); and their combinations with `ws`. The hard commit `c2hard`, the `c2` generation truncated where the forced box closes and scored as a natural stop, and the replications on two more students were added after the first numbers. Table 36 gives T=6,144.

- **Placebo and suppression do nothing.** `resume` moves capping and accuracy by at most 0.5 points (every CI contains zero), and `ws` alone moves wrong-and-capped by -0.6 to $+0.4$: suppressing the cues of re-derivation does not produce a stop.
- **A commit inside the thought is half a commit.** `c1` cuts wrong-and-capped by 3.6 points and raises solved-then-capped from 2.3 to 4.5%: the student boxes an answer and keeps thinking.
- **The soft commit halves the branch.** `c2ws` takes wrong-and-capped from 12.7 to 6.3% and raises accuracy by 3.2 and delivered accuracy by 2.7 points, turning 2 of 834 solved generations wrong against the placebo's 9. The arm the registered rule picks, `c2ws` at T=4,096, changes the branch by -6.5 points (95% CI $[-8.6, -4.5]$), removing 51% of it; the prediction asked for 6.6, half of the 13.2-point branch it assumed, and misses by 0.1 because this cell's branch is 12.7.
- **Stuck states hold an answer a quarter of the time.** Forced answers are right on 25–29% of the stuck generations whose untouched run ended wrong ($31/124$, $31/115$, $32/111$ at the three T): the state mostly holds no correct answer, and the quarter that does is the accuracy gain.
- **The hard commit removes most of it.** `c2hard` takes wrong-and-capped from 12.7 to 1.7% (87% of the branch) and capping from 15.0 to 3.1%, with accuracy $+2.6$, delivered $+3.5$ and 5 of 834 solved generations turned wrong. Under NO-SYSTEM it takes the branch from 13.2 to 2.6% (6.9% soft), and the format gap in capping changes by $+0.4$ (-0.4 soft), not significant either way.
- **It replicates.** On the seed-123 student (one decoding seed) the soft commit cuts wrong-and-capped by 6.0 points and the hard one by 12.0 (accuracy $+3.2$, $+2.8$); on the Qwen2.5-7B student by 11.6 and 15.8 (accuracy $+3.7$, $+3.9$). On the base, 7 of 500 generations are stuck and accuracy moves by $+0.2$, not significant.

The soft commit leaves the stop to the student, and its stop is answer-conditioned: after the forced close and answer it ends the turn 84% ($48/57$) of the time when the forced answer is right and 36% ($38/107$) when it is wrong (`c2`, T=4,096, Fisher $p=1.5\times10^{-9}$), in the same direction at every T and on both replication students, while the base stops after 83–100% of forced wrong answers (Table 37). Of the 63 wrong-and-capped generations left under `c2ws` at T=6,144, 50 were triggered and capped again, most of them writing on at length after the forced answer. Rule (b) is therefore a hard commit: the student has 64 tokens to write its answer and the turn then ends. Rule (a) handles the generations the trigger never touches, those already holding an answer, which is where the format gap's solved

branch sits. The right-versus-wrong contrast is observational and may partly reflect confidence or item difficulty; the hard commit delivers a bare boxed answer without an explanation; the study used two decoding seeds (one for the seed-123 student and the base).

Table 36: **A commit on stuck generations halves the wrong branch when the student may keep going and removes most of it when the answer ends the turn; the placebo and *Wait*-suppression do nothing.** MATH-500, budget 8,192, trigger at T=6,144; *stuck*: generations triggered (of 1,000; 500 for the seed-123 student and the base); w&c: wrong-and-capped %; Δ: paired change from untouched; stop-after-wrong in %; solved→wrong: untouched-correct generations the arm makes wrong; § added after the first numbers; * significant.

arm	stuck	cap	w&c	Δ w&c	acc.	Δ acc.	stop-after-wrong	solved→wrong
main student, EXACT, untouched		15.0	12.7		83.4		23.5	
`resume` placebo	130	15.1	12.8	+0.1	83.3	−0.1	23.4	9/834
`ws` suppress *Wait*	130	15.1	12.7	0.0	83.5	+0.1	23.0	9/834
`c1` commit in thought	130	13.6	9.1	−3.6*	86.6	+3.2*	32.1	3/834
`c1ws` c1 + ws	130	13.7	9.1	−3.6*	86.6	+3.2*	32.1	1/834
`c2` close, answer	130	9.4	6.9	−5.8*	86.3	+2.9*	49.6	5/834
`c2ws` c2 + ws	130	9.1	6.3	−6.4*	86.6	+3.2*	53.0	2/834
`c2hard` c2, stop at brace§	130	3.1	1.7	−11.0*	86.0	+2.6*	87.9	5/834
main student, NO-SYSTEM, untouched		31.6	13.2		82.8		23.3	
`resume` placebo	124	31.5	13.1	−0.1	82.9	+0.1	23.4	4/828
`c2ws` c2 + ws	124	25.3	6.9	−6.3*	86.3	+3.5*	49.6	1/828
`c2hard` c2, stop at brace§	124	20.1	2.6	−10.6*	86.0	+3.2*	81.4	2/828
seed-123 student, EXACT, untouched		16.6	13.8		81.0		27.4	
`resume` placebo	66	16.2	13.6	−0.2	80.8	−0.2	29.2	2/405
`c2ws` c2 + ws	66	11.6	7.8	−6.0*	84.2	+3.2*	50.6	0/405
`c2hard` c2, stop at brace§	66	4.0	1.8	−12.0*	83.8	+2.8*	88.9	0/405
Qwen2.5-7B student, EXACT, untouched		19.9	18.2		74.9		27.5	
`resume` placebo	170	20.1	18.4	+0.2	74.8	−0.1	27.0	5/749
`c2ws` c2 + ws	170	8.0	6.6	−11.6*	78.6	+3.7*	69.2	1/749
`c2hard` c2, stop at brace§	170	3.5	2.4	−15.8*	78.8	+3.9*	88.7	1/749
Qwen3-8B-Base, EXACT, untouched		1.8	1.8		55.2		96.0	
`resume` placebo	7	2.0	2.0	+0.2	55.2	0.0	95.5	0/276
`c2ws` c2 + ws	7	0.8	0.8	−1.0*	55.4	+0.2	98.2	0/276
`c2hard` c2, stop at brace§	7	0.8	0.8	−1.0*	55.4	+0.2	98.2	0/276

Table 37: **After a forced answer the students end the turn mostly when it is right; the base ends it either way.** Natural stops after the `c2` commit among triggered generations, split by whether the forced answer is right or wrong (stopped/total), EXACT, with Fisher's exact p.

	T=4,096			T=6,144			T=7,168		
student	right	wrong	p	right	wrong	p	right	wrong	p
main student	48/57	38/107	1.5×10^{-9}	32/41	31/89	4.6×10^{-6}	27/40	28/82	8.7×10^{-4}
seed-123 student	18/23	18/61	1.3×10^{-4}	12/18	16/48	0.02	7/14	11/47	0.09
Qwen2.5-7B student	56/63	94/146	2.1×10^{-4}	43/47	81/123	4.8×10^{-4}	24/40	48/116	0.05
Qwen3-8B-Base	1/1	6/7	1	1/1	5/6	1	1/1	5/6	1

A settledness gate (exploratory). On 17,149 check-point states from 3,000 of the student's MATH-training rollouts we labelled each state by the value of continuing (forced answer now against the rollout's own outcome) and fit logistic gates. Read from the last-layer hidden state the gate barely beats elapsed position (held-out AUC 0.653 against 0.640) and, replayed on MATH-500, ties the decode guard at matched tokens (−0.2 [−0.7, +0.3]; pre-registered, failed). Adding settledness features—whether the forced answer at a check point equals the previous one, and how many distinct answers it has held—raises the AUC to 0.825. A rule that needs no labels, committing once three consecutive forced answers (every 512 tokens) agree, delivers 84.0% at a mean 1,747 tokens against 82.3% for the decode guard at matched tokens (+1.7 [+0.6, +2.8]) with fewer premature commits (16 against 23); the token counts exclude the up-to-64-token probe answers the rule computes at each check point. An oracle stop time with the same commit action would reach 88.7% at 2,367 tokens.

A12 PROTOCOL: PROMPTS, METRICS AND TRAINING

A12.1 THE EXACT PROMPTS

All prompts end immediately before the first generated token; no <think> is pre-filled. Each \n is one newline byte; line breaks inside the listings are layout only and carry no byte. The removed system turn is 11 Qwen3 tokens (<|im_start|>, system, \n, You, are, a, helpful, assistant, ., <|im_end|>, \n); an empty one is 5.

```
EXACT:      <|im_start|>system\nYou are a helpful assistant.<|im_end|>\n
            <|im_start|>user\n{problem}<|im_end|>\n<|im_start|>assistant\n
NO-SYSTEM:  <|im_start|>user\n{problem}<|im_end|>\n<|im_start|>assistant\n
INSTR:      ... user\nSolve the following math problem. Make sure to put the answer
            (and only answer) inside \boxed{}.\n\n{problem}<|im_end|> ...
HISTORICAL: NO-SYSTEM with the INSTR user turn.
ALT-SYSTEM: system text replaced by "Please reason step by step, and put your final
            answer within \boxed{}."
```

Stop tokens: <|im_end|> (151645) and <|endoftext|> (151643).

```
EXACT:      <|begin_of_sentence|>You are a helpful assistant.<|User|>{problem}<|Assistant|>
NO-SYSTEM:  <|begin_of_sentence|><|User|>{problem}<|Assistant|>
```

Stop token: <|end_of_sentence|> (151643).

Llama-3.1 template (Llama-3.1-8B-Instruct and Llama-3.1-8B base students). The trainer's template injects a date preamble into the system turn—the training day—which we reproduce byte-for-byte (02 May 2026 for the Instruct student, 18 Sep 2026 for the base student, read back from each run's logged tokenized example):

```
EXACT:      <|begin_of_text|><|start_header_id|>system<|end_header_id|>\n\n
            Cutting Knowledge Date: December 2023\nToday Date: 02 May 2026\n\n
            You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>\n\n
            {problem}<|eot_id|><|start_header_id|>assistant<|end_header_id|>\n\n
NO-SYSTEM:  the same without the system block.
```

Stop tokens: <|eot_id|> (128009) and <|end_of_text|> (128001).

Qwen2.5-7B base student. The ChatML format above (Qwen3 template), byte-identical.

Mistral template (Mistral-7B-v0.3 base student). The trainer's mistral_nemo template places the system text inside the instruction block; there is no structural system slot:

```
EXACT:      <s>[INST] You are a helpful assistant.\n\n{problem}[/INST]
NO-SYSTEM:  <s>[INST] {problem}[/INST]
```

Stop token: </s> (2).

A12.2 OUTCOME DEFINITIONS

A budget that shrinks without saying so. One replication hazard is worth naming, because it would counterfeit the effect this paper measures. Our runner sets the generation budget as $\max(16, \min(B,\ L - |p| - 4))$, where B is the nominal budget, L the model's context length and $|p|$ the prompt length. If L is set too small the run does not fail; it silently caps generations below B, and a generation truncated that way is indistinguishable in the output from one that ran to the budget. Anyone repeating these experiments with a shorter context would therefore manufacture our result. We checked every cell rather than assuming: tokenising all GPQA and LiveCodeBench prompts under both formats with each student's own tokenizer, the longest prompt is $2{,}803$ tokens, so L must be at least $27{,}383$ for a $24{,}576$-token budget; at the $32{,}768$ we use, zero items in any cell have their budget reduced. *cap*: the generation reached the token budget. *accuracy*: the

last \boxed{...} in the generation is equivalent to the reference, judged by math-verify symbolic equivalence with a string-normalization fallback (the string-only judge gives 9–13 points lower absolute accuracies on MATH-500, near-identical ones on AIME, and identical conclusions). *solved-then-capped*: cap and accuracy both true. *loop*: over the last 3,000 whitespace tokens, $1 - \#\text{distinct 20-grams}/\#\text{20-grams} > 0.5$. *well-formed stop*: not capped and the text contains </think> or <|end_of_thought|> (both markers are accepted; released OpenThoughts-style students close with the second). *premature stop*: not capped and no \boxed{} anywhere. *stop-when-wrong*: among generations whose last boxed answer is not equivalent to the reference (or that have none), the share that stopped naturally. Comparisons are paired by (problem, seed); 95% CIs resample problems (item-clustered bootstrap, 5,000 draws), and a difference is significant when its CI excludes zero.

A12.3 Training details

Generations come from two installations of the same vLLM version (a local node and cluster nodes); paired comparisons never cross installations, and cells reported from both agree to within about one point (MATH-500 EXACT cap 15.0 vs 15.7%, NO-SYSTEM 31.6 vs 30.9%). Original SFT: full fine-tuning, 3 epochs over 30,000 traces, packing to 32,768 tokens, peak learning rate 5×10^{-6} with 3% warm-up and cosine decay, effective batch 32 packed sequences (8 GPUs $\times$ 1 $\times$ 4 accumulation steps), AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.1, gradient clipping 1.0, bf16, ZeRO-3 with CPU offload, flash attention. Every locally trained student uses this recipe on one of six initializations (Qwen3-8B-Base, Qwen2.5-7B, Llama-3.1-8B base and Instruct, Mistral-7B-v0.3, and the 32B, the last with context parallelism), except where stated: the one-epoch and 100k-problem students (Appendix Table 67) change the epochs and data, the science and code students the data (and, for code, the teacher; Appendix A12.4), and the optimizer and sharpened runs only β_1 or the loss (Appendix A15.1). Repair arms: one epoch over 3,420 rows starting from the CE checkpoint, learning rate 2×10^{-6} constant after 5% warm-up, effective batch 16 packed sequences, otherwise identical trainer settings. Arms A and B supervise only the closing segment (from </think> to the end of turn) of their 1,710 non-replay rows and use full loss on the 1,710 replay rows; arms C and D use full loss on every row. In arm D the 912 perturbed rows are chosen at random and formatted under NO-SYSTEM or HISTORICAL with equal probability; the assistant target is unchanged. Arms B, C, D (seed 42) were trained on a local node (DeepSpeed 0.16.7, torch 2.6) and the corrected arm A, the two full-loss ablations, and the seed-123 and 32B arms on cluster nodes (DeepSpeed 0.19.6, torch 2.8) with identical trainer arguments (two GPUs $\times$ 8 accumulation steps for the cluster arms, same effective batch); all cells of Table 77 are decoded on the cluster stack against a CE baseline decoded there, and the A vs. A-full contrast is within one training stack while B vs. B-full spans the two.

The first arm A was trained on corrupted rows. In the first build of arm A the prompt written into each row came from a different index than the rollout it was paired with for 912 of its 1,710 non-replay rows, and 895 rows lost the <think> opener, so the trainer inserted an empty thought before the supervised closing segment. We found this while auditing the arm's inputs after the paper's first draft. The build was repaired (the prompts reformatted from the rollout's own problem, the openers restored; 1,710 rows, merged file SHA-256 43d17a50...4d11b) and the arm retrained from the same checkpoint with the same recipe. Every arm A number in this paper is from the corrected run; the original arm A looked worse—it raised loops under ALT-SYSTEM from 1.4 to 8.6%—and that apparent side effect does not survive the repair.

A12.4 Training data, decontamination and harness defaults

The three training sets differ in more than domain (Table 38a): mathematics and code are long-CoT (median thought 7.8k and 9.0k tokens), the science set has short thoughts followed by long markdown answers, supplies about three times fewer tokens per epoch and therefore gave 267 optimizer steps against 855 and 936 under identical hyper-parameters. The code traces open mid-sentence and are never boxed. Problem texts were decontaminated against all four evaluation sets (Table 38b): no exact matches, no evaluation item with more than 43% of its 13-grams in one training prompt, and, from 8-gram candidates read by hand, three MATH-500 problems (0.6%; four in the 100k pool)

present as reformatted near-copies in the math training set; none for AIME 2025, GPQA-Diamond or LiveCodeBench.

Table 38: (a) Training sets: median [p10, p90] over all rows, Qwen3 tokens on the training format; "kept" = rows of at most 32,768 tokens (longer rows are dropped by the trainer, not truncated). †The source does not document its generator. §Gemini Pro 85.6%, Flash 14.4%. (b) Exact matches / items sharing a non-boilerplate 13-gram with some training prompt / largest fraction of an item's distinct 13-grams in one training prompt / near-verbatim duplicates confirmed by reading every 8-gram candidate with word-sequence similarity ≥ 0.6. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

(a) training sets	math	science	coding
teacher	gpt-oss-120b	undocumented†	Gemini Pro/Flash§
rows / kept (≤32k tok.)	30,000 / 29,991	30,000 / 30,000	29,884 / 29,834
thought tokens	7,826 [2,644, 16,600]	371 [142, 1,589]	8,984 [2,977, 17,963]
answer tokens	882 [481, 1,333]	2,160 [978, 3,914]	477 [167, 1,199]
`\boxed` in answer	100.0%	38.6%	0.0%
top thought opening	"We need to" 78.5%	"We need to" 63.1%	"that leads to" 63.3%
tokens / epoch	297.5M	92.7M	325.7M
packed 32k seqs / epoch	9,115	2,851	9,975
optimizer steps (3 ep.)	855	267	936

(b) decontam.‡	MATH-500 (500)	AIME 2025 (30)	GPQA-D (198)	LCB (182)
math 30k	0 / 14 / .43 / 3	0 / 9 / .15 / 0	0 / 0 / – / 0	0 / 3 / .03 / 0
math 100k (pool)	0 / 17 / .43 / 4	0 / 9 / .27 / 0	0 / 0 / – / 0	0 / 4 / .03 / 0
science 30k	0 / 0 / – / 0	0 / 0 / – / 0	0 / 0 / – / 0	0 / 0 / – / 0
coding 30k	0 / 0 / – / 0	0 / 0 / – / 0	0 / 0 / – / 0	0 / 99 / .08 / 0

How much a harness default moves the number. With the weights, the problems and the decoder held fixed and only the prompt bytes differing, the four chat-template harness defaults of Table 40 report the seed-42 student's MATH-500 behaviour as follows (n=500 per cell, one decoding seed, budget 8,192; all six cells generated on the same cluster, since per-item outcomes do not reproduce across hardware and pairing across clusters is therefore invalid).

prompt source	cap	loop	well-formed	accuracy
its own training format	15.6	1.0	84.4	83.0
NeMo-Skills	25.4	13.4	57.2	81.2
OpenCompass / EvalScope	26.4	9.4	66.8	80.6
lighteval	19.6	7.0	72.2	77.8
simple-evals	13.0	1.4	87.0	79.4
range across harness defaults	13.4	12.0	29.8	4.4

Capping spans 13.4 points and well-formed stops 29.8, against the run-to-run noise of about one point established above. Two further defaults widen it, and they widen it for a different reason, which is why we keep them in a separate sentence. lm-evaluation-harness applies no chat template at all. Under its `hendrycks_math500` task the same checkpoint caps 51.4% of generations, and under the 4-shot `minerva_math500` task it caps 91.0% with accuracy falling to 10.0%—on a student that answers 83.0% correctly and caps 15.6% under the format it was trained on. Every one of these cells ran at our 8,192-token budget rather than lm-evaluation-harness's own 256, so this is not a budget effect. Across all six task configurations the reported capping rate spans 13.0 to 91.0%. That 78-point figure is real but it measures something coarser than the rest of this paper: serving a chat-tuned student as a raw completion, with few-shot exemplars, does not perturb its prompt format so much as hand it to a different model. The 13.4-point range across chat-template defaults is the one that bears on our argument, because it is what two reasonable people running two mainstream harnesses on the same checkpoint actually get. One number from that cell must not be read at face value: its well-formed-stop rate is 0.0% by construction, because a student given a raw completion prompt never opens the thinking block and so cannot close one; 49% of those generations do stop naturally, just without the marker. The one-sentence version is that the same checkpoint on the same benchmark reports 13% or 26% budget-capped generations depending only on which mainstream harness runs it, and 51% if that harness declines to apply a chat template. An independently trained student—the second training seed, same recipe—reproduces this: its capping spans 13.8 to 36.8% across the same four chat-template defaults, a 23.0-point spread, and 47.4% under lm-evaluation-harness. The ordering of harnesses is broadly preserved between the two students, with simple-evals gentlest and lm-evaluation-harness harshest in both.

The obvious question is whether every model swings this much across prompts, and the answer is no. Run through the identical set of defaults, DeepSeek-R1-Distill-Qwen-7B—a released student post-trained under several serializations—spans 1.8 to 4.2% of capping, a range of 2.4 points, against our two students' 13.4 and 23.0. The spread is a property of the student, not of evaluation harnesses in general. A harness default is a large perturbation only for a model that was fine-tuned under one format.

A released checkpoint, a public harness, no prompt of ours. The released panel of Appendix A15.4 contains one positive: TinyR1-32B-Preview gains +11.0 points of capping when a generic system turn is *added*, where its parent DeepSeek-R1-Distill-Qwen-32B is flat under the byte-identical edit. That edit was still ours. Of the four chat-template defaults exactly one, simple-evals, sends a system message ("You are a helpful assistant."); TinyR1's DeepSeek template has no system slot, so that message lands as bare text between the BOS and the user marker—the same structural edit, authored by a public harness. We pre-registered three predictions before any of these cells existed: (P1) TinyR1 spans more than 5 points across the four defaults with simple-evals its harshest column; (P2) its parent spans less than 5 and simple-evals is not its harshest; (P3) Qwen3-4B-Thinking-2507, Qwen3-14B and Light-R1-14B-DS each span less than 5. All three held (Table 7): TinyR1 caps on 33.2% under simple-evals against 1.0 under lighteval, a spread of 32.2 points—the largest in the panel, twice our own students'—with detected loops spanning 35.4 points; its parent spans 1.2, the other three 4.6, 1.4 and 1.8. The same released model reads as essentially never capping or as capping on a third of the benchmark depending only on which mainstream harness runs it, and the column that breaks it is the one that inserts a system turn.

One number from that row must be suppressed rather than reported. Its accuracy appears to span 86.4 points across those columns, which is entirely an artefact of the answer protocol: DeepSeek complies with lighteval's `ANSWER:` instruction and stops boxing, so a boxed-answer judge reads it at 3.0% there. Any accuracy range spanning columns whose requested answer format differs is meaningless, and we do not quote one. This is the second metric with that property, after well-formed stops under raw-completion prompts; capping, looping and natural stops need no answer extraction, which is why the headline rests on them. Paired against the NeMo-Skills column, lighteval differs by −5.8 points of cap and simple-evals by −12.4, both significant, while OpenCompass does not differ significantly. The harness is not a uniform penalty: simple-evals' default leaves this student *better* terminated than its own training format does, and OpenCompass's leaves it nearly twice as badly. A headline termination number therefore reports a property of the pair (model, harness default), not of the model.

A pre-registered census of three further released checkpoints. To test whether the recipe label alone predicts harness spread we pre-registered, before running, that single-stage single-dataset distills move (> 5 points across the four defaults) and multi-stage or multi-format checkpoints do not, then ran three that fit on the cluster (nine more were pre-registered and cut for contention before any result was read; their predictions stand in the file). The two multi-stage checkpoints are flat as predicted: QwQ-32B spans 1.0 points and OpenReasoning-Nemotron-32B 0.6. The single-stage one, s1.1-32B, was predicted to move and did not: 2.0 [0.0, 4.0]. We report that as a failed prediction, with the reading the panel itself supplies: s1.1's template injects the Qwen default system turn in every harness column, so the four defaults never differ in the one thing the account says matters, and it is fine-tuned from an instruct checkpoint. Re-reading the panel by that criterion, the released checkpoints whose served template carries a system turn that some default removes (TinyR1-32B-Preview, DeepSeek-R1-0528-Qwen3-8B) are the movers, and every row in which all four defaults leave the system-turn state unchanged is flat, whatever its recipe. The recipe label is not the predictor; whether the harness can take the trained format away is. (s1.1's well-formed column reads 0.0 by construction—its closer is not among the markers the scorer recognises—and is not a result.) Existing rows are byte-identical before and after the census; receipts in the release.

The same effect in accuracy, and it inverts between models. Two of these harnesses ask for the answer on a line beginning `ANSWER:` rather than in a box, and whether a model obeys that instruction decides its reported accuracy far more than its mathematics does. Under lighteval's default our seed-42 student emits the requested line in 3.4% of generations and keeps boxing in 83.2%; DeepSeek-R1-Distill-7B does the opposite, complying in 90.0% and boxing in 4.2%. Scored by a boxed-answer judge the two read 77.8 and 3.0; scored by lighteval's own protocol they read 2.4 and

66.8. Same harness, same judge within each column, and the ranking of the two models reverses. Neither pair of numbers measures mathematical ability. Our termination metrics are unaffected—capping, looping and well-formed stops need no answer extraction, which is why the range above rests on them—but it does mean the accuracy column of a cross-harness comparison is not interpretable unless the judge matches the protocol the prompt asked for.

One further caution. These cells are not interchangeable across hardware: on two clusters the same prompt gives aggregates within about a point but per-item agreement of only 79.8% on capping and 64.6% on well-formed stops, which is enough to destroy a paired comparison while leaving the means looking reproducible. Every paired comparison in this paper is made within one cluster.

Table 39 lists what public harnesses produce by default. The math and science defaults of NeMo-Skills, lm-evaluation-harness, lighteval, OpenCompass and EvalScope insert no system turn (only NeMo-Skills reproduces our HISTORICAL format exactly; lm-evaluation-harness applies no chat template unless asked); the official LiveCodeBench runner and EvalScope's code task do insert one. The Qwen2.5 chat template re-inserts "You are a helpful assistant." when the caller gives none; the Qwen3, DeepSeek-R1-Distill and our Llama-3.1-Instruct templates do not.

Table 39: Default prompt format of public evaluation harnesses (pinned commits) for chat models. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

harness (commit)	default system turn	default instruction line (user turn)
NeMo-Skills (`e80d524`)	none	math: our HISTORICAL line; code: reasoning preamble inside the user turn
lm-evaluation-harness (`d6de816`)	none; no chat template unless flagged	AIME: "Question: … Answer:"
lighteval (`6ba40c4`)	none	AIME: "Solve the following math problem efficiently…"
OpenCompass (`a2117d4`)	none	AIME: "… Remember to put your final answer within \boxed{}." (after the question)
EvalScope (`06b4c67`)	AIME: none; LiveCodeBench: "You are an expert Python programmer…"	AIME: "Solve the following math problem step by step…"
LiveCodeBench official (`28fef95`)	Qwen2/2.5 instruct entries: "You are a helpful assistant."	problem-specification preamble
simple-evals (`652c89d`)	GPT chat: "You are a helpful assistant."; o-series: none	"Solve the following math problem step by step…"

Table 40: **What seven public harnesses send for MATH-500 by default**, re-derived programmatically from each harness's pinned checkout (`harness_default_prompts_v1.py`; all seven literals machine-checked against source, and the NeMo-Skills row additionally verified by executing the library's own prompt builder). *system turn* is the system message the harness puts in the request, not necessarily what the model sees: a chat template may insert one when the caller supplies none (OpenThinker-7B's Qwen2.5-Instruct template inserts "You are Qwen, created by Alibaba Cloud. You are a helpful assistant.", OpenMath-Nemotron-7B's inserts an *empty* system block, our Qwen2.5-7B students' inserts "You are a helpful assistant.", and the Qwen3 and DeepSeek-R1-Distill templates insert nothing). *answer protocol* is the form the harness's own grader reads; two harnesses ask for a line beginning `ANSWER:`/`Answer:` and simple-evals explicitly instructs the model *not* to use `\boxed`, so a `\boxed`-only judge under-reports their accuracy. lm-evaluation-harness additionally differs by more than prompt bytes: it applies no chat template, decodes greedily, stops at `"Problem:"`, and its vLLM backend defaults to `max_gen_toks`= 256. LiveCodeBench (`28fef95`) has no MATH-500 task and is omitted.

harness (commit)	MATH-500 task	system turn	default user turn (P = problem text; `\n` = newline)	answer protocol
NeMo-Skills (`e80d524`)	`math-500`	none	`Solve the following math problem. Make sure to put the answer (and only answer) inside \boxed{}.\n\n`P	`\boxed{}`
lm-evaluation-harness (`d6de816`)	`hendrycks_math500`	*no chat template*	`Problem:` P`\nAnswer:`	`\boxed{}`
lm-evaluation-harness (`d6de816`)	`minerva_math500`	*no chat template*	4-shot: four `Problem:\n`P_i`\n\nSolution:` + solution blocks (target delim `" "`, shot delim `"\n\n"`), then `Problem:\n`P`\n\nSolution:`	`\boxed{}`
lighteval (`6ba40c4`)	`math_500`	none	`Solve the following problem. The final line of your response MUST be of the following format:\n"ANSWER: $ANSWER" (without quotes) where $ANSWER is the final answer. Think step by step before answering.\n\n`P	`ANSWER:`
OpenCompass (`a2117d4`)	`math-500`	none	P`\nPlease reason step by step, and put your final answer within \boxed{}.`	`\boxed{}`
EvalScope (`06b4c67`)	`math_500`	none	*byte-identical to the OpenCompass row*	`\boxed{}`
simple-evals (`652c89d`)	`math_500_test`	`You are a helpful assistant.`	`Solve the following math problem step by step. The last line of your response should be of the form Answer: $ANSWER (without quotes) where $ANSWER is the answer to the problem.\n\n`P`\n\nRemember to put your answer on its own line after "Answer:", and you do not need to use a \boxed command.`	`Answer:`

Independently of the byte check, that format also reproduces in *outcome*. Regenerating the same prompt on different hardware and a different vLLM build gives cap 25.4 against 24.8%, loops 13.4

against 12.0, well-formed stops 57.2 against 56.6 and accuracy 81.2 against 79.8—no difference significant at n=500. That fixes a scale for run-to-run noise in this pipeline at roughly one point on a single-arm rate. The paired differences this paper reports are quieter still, because noise common to the two arms cancels: two released models happened to complete on both clusters, and their paired change in capping reproduces to within 0.2 and 0.4 points respectively, with the single-arm rates reproducing to within 0.6 and 0.8. Those are the yardsticks every number here should be read against. The effects we report are tens of points and the differences we decline to interpret are two or three, which is a few multiples of the paired noise floor rather than a hair above it.

The HISTORICAL format is not our reconstruction of what NeMo-Skills sends: it is byte-identical to the prompt that harness's own library produces, checked by executing `get_prompt('generic/math').fill` at the pinned commit. Under it the seed-42 student caps 24.8% of MATH-500 generations against 15.6% under the format it was trained on, a significant paired difference of -9.2 points in favour of the training format. Two further cautions follow from Table 40 and apply to any cross-harness comparison, including ours. Two of the seven harnesses ask for an `ANSWER:` line rather than a boxed answer, and simple-evals instructs the model *not* to use `\boxed` at all, so a boxed-answer judge scores those configurations near zero for reasons that have nothing to do with the model. And lm-evaluation-harness differs from the others by more than prompt bytes—no chat template, greedy decoding, a `"Problem:"` stop string, and a vLLM-backend default of 256 generated tokens—so its numbers belong on a separate axis from the prompt-format one and we do not pool them.

A12.5 Robustness of the loop metric

The loop rate is a repetition statistic (20-gram repetition ratio >0.5 over the last 3,000 whitespace tokens). Recomputing it at thresholds 0.3 / 0.7 and windows 1,500 / 6,000 whitespace tokens, and replacing it by two independent detectors—a periodicity test on whitespace tokens (smallest period $p \leq 400$ for which $\geq 90\%$ of the last 800 satisfy $w_i = w_{i-p}$) and an exact repeated 50-token substring of tokenizer tokens occurring at least three times in the last 3,000 tokenizer tokens—gives Cohen's κ of 0.63–0.98 with the paper's definition on every CE cell (≥ 0.82 except the substring detector on MATH-500 EXACT and AIME HISTORICAL) (MATH-500 EXACT / NO-SYSTEM, n=1000; AIME EXACT / HISTORICAL / NO-SYSTEM, n=360), and the EXACT→NO-SYSTEM gap is the same under every definition (MATH-500: 1.3 → 15.1 paper, 1.6 → 15.7 periodic, 2.8 → 19.3 substring; AIME: 11.1 → 25.0, 11.9 → 25.8, 15.3 → 32.2). Of the generations the paper calls loops, 97–100% are strictly periodic under the independent test and 100% are budget-capped; the metric is conservative rather than permissive.

A12.6 What the training format leaves, and fidelity to a standard harness

Table 41: Base model, long-CoT students and a no-think student under the two formats. Each cell: cap / accuracy / stop-when-wrong (%), where stop-when-wrong is the share of incorrect generations that stopped naturally. MATH-500 one decoding seed (n=500); AIME four seeds (n=120). "no-think" students were not evaluated on AIME. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

	MATH-500		AIME 2025	
student	EXACT	HISTORICAL	EXACT	HISTORICAL
Qwen3-8B-Base (no SFT)	2.0 / 53.2 / 95.7	–	6.7 / 10.0 / 92.6	–
CE long-CoT student	15.6 / 83.0 / 22.4	24.8 / 79.8 / 34.7	52.5 / 29.2 / 29.4	62.5 / 21.7 / 25.5
no-think student (same problems, answers only)	0.0 / 68.0 / 100.0	0.0 / 64.8 / 100.0	–	–
long-CoT student, independent re-run	14.6 / 81.6 / 25.0	31.2 / 80.4 / 28.6	43.3 / 28.3 / 39.5	72.5 / 18.3 / 24.5

Table 41 sets the long-CoT students beside their base and a no-think student trained on the same problems, which never caps on MATH-500. On checkpoints that had previously been evaluated with the NeMo-Skills harness, our pipeline under the HISTORICAL format reproduces that harness's numbers within seed noise (e.g. AIME cap 58.3 vs 60.6% historical for the CE student, both n=360; MATH-500 cap 31.2 vs 34.0% and loop 17.6 vs 20.2% for the eight-epoch student), which indicates that the difference between EXACT and HISTORICAL is a property of the format and not of the decoding stack.

A13 THE EFFECT, REPLICATED AND STRESS-TESTED

A13.1 SEVEN STUDENTS UNDER THE TWO PROMPT FORMATS

Table 42: Qwen3-8B-Base CE student, only the prompt format changes. MATH-500: budget 8,192, one decoding seed, n=500 per cell. AIME 2025: budget 30,000, twelve decoding seeds, n=360. Columns: cap, accuracy, solved-then-capped, loop, well-formed stop (%), median tokens. The remainder to 100 after cap and well-formed are unclosed stops; AIME medians are censored at the budget. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	cap	acc	solved-then-capped	loop	well-formed	med. tokens
EXACT (training format)	15.6	83.0	2.4	1.0	84.4	1,549
system + instruction line	13.8	84.6	2.2	0.6	86.2	1,228
HISTORICAL (no system + instruction line)	24.8	79.8	11.6	12.0	56.6	1,202
NO-SYSTEM (no system, bare problem)	33.0	82.4	19.6	16.6	54.0	1,909
AIME 2025, EXACT (n=360)	49.7	26.1	1.1	11.1	50.3	29,638
AIME 2025, system + instruction line (n=360)	50.0	26.7	1.1	14.4	50.0	30,000
AIME 2025, HISTORICAL (n=360)	58.3	26.4	6.1	18.3	32.2	30,000
AIME 2025, NO-SYSTEM (n=360)	67.2	24.4	8.6	25.0	21.7	30,000
AIME 2025, ALT-SYSTEM (held-out system prompt) (n=360)	53.9	26.1	1.4	12.2	45.8	30,000
AIME 2025, EMPTY-SYSTEM (n=360)	53.3	25.3	1.1	12.8	40.3	30,000

Table 43: Removing the training system turn, seven students (six initializations) trained with the same recipe, and the DeepSeek-32B parent before our SFT, MATH-500, budget 8,192, one decoding seed, n=500 per cell. The Llama students' format includes the Llama-3.1 date preamble that the trainer injected; the Mistral format is `<s>[INST] {system}\n\n{problem}[/INST]` (Appendix A12.1). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student (initialization)	prompt format	cap	acc	solved-then-capped	loop	well-formed
Qwen3-8B-Base	EXACT	15.6	83.0	2.4	1.0	84.4
	NO-SYSTEM	33.0	82.4	19.6	16.6	54.0
	HISTORICAL	24.8	79.8	11.6	12.0	56.6
DS-R1-Distill-Qwen-32B	EXACT	10.0	92.2	4.8	1.2	90.0
	NO-SYSTEM	49.6	79.4	30.0	17.4	50.4
	HISTORICAL	14.2	91.2	7.8	4.6	85.8
DS-R1-Distill-Qwen-32B, no local SFT	EXACT	5.2	85.8	0.0	0.0	94.8
	NO-SYSTEM	7.4	89.8	1.0	0.0	92.6
	EXACT, pooled (n=1000)	5.4	85.7	0.4	0.1	94.6
	NO-SYSTEM, pooled (n=1000)	7.1	89.9	0.9	0.0	92.9
Llama-3.1-8B-Instruct	EXACT	28.6	61.2	2.0	1.2	71.4
	NO-SYSTEM	28.8	60.6	3.8	2.6	62.6
Llama-3.1-8B (base)	EXACT	33.6	48.2	2.0	4.8	66.4
	NO-SYSTEM	36.8	45.8	4.4	6.0	63.2
	EXACT, pooled (n=1000)	35.8	47.2	2.1	5.1	64.2
	NO-SYSTEM, pooled (n=1000)	35.8	46.5	3.1	6.1	64.2
Qwen2.5-7B (base)	EXACT	21.6	75.0	3.0	1.6	78.4
	NO-SYSTEM	58.0	74.8	37.2	24.2	41.8
	EXACT, pooled (n=1000)	19.9	74.9	1.7	1.6	80.1
	NO-SYSTEM, pooled (n=1000)	59.9	76.6	39.4	28.3	40.0
	HISTORICAL	70.6	74.8	47.8	41.6	29.4
Qwen2.5-7B (base), seed 123	EXACT	20.4	74.6	2.2	0.6	79.6
	NO-SYSTEM	51.2	75.6	31.2	21.2	48.8
	EXACT, pooled (n=1000)	20.9	74.1	2.0	1.3	79.1
	NO-SYSTEM, pooled (n=1000)	51.9	74.9	31.9	21.7	48.1
Mistral-7B-v0.3 (base)	EXACT	36.2	52.8	3.8	2.6	63.8
	NO-SYSTEM	38.8	50.8	3.8	3.0	61.2
	EXACT, pooled (n=1000)	37.2	51.4	3.3	3.0	62.8
	NO-SYSTEM, pooled (n=1000)	37.9	49.9	3.0	3.6	62.1

A13.2 THE SAME FAILURE THROUGH AN ORDINARY SERVING ENDPOINT

Every other number in this paper comes from our own evaluation harness, so the perturbation could be read as one we authored. This appendix removes that reading. We served the seed-42 student behind vLLM's OpenAI-compatible endpoint with the checkpoint's own

chat template and vanilla API defaults, and sent the 500 MATH-500 problems twice: once as `messages=[{"role":"user"}]`, which is what a caller who sends only a question produces, and once with the default system message prepended. Two decoding seeds, budget 8,192, temperature 0.6, top-p 0.95, n=1000 per arm, paired by (problem, seed).

Table 44: The seed-42 student behind a standard serving endpoint. Percentages of generations; the last row is the paired difference. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

	cap	accuracy	solved-then-capped	loop	well-formed
system message present	15.7	83.1	3.0	1.3	84.3
user message only	32.9	80.0	16.5	15.7	54.8
paired difference	$+17.2^*$	-3.1^*	$+13.5^*$	$+14.4^*$	-29.5^*

Two things make this more than a replication. First, the prompt is not ours. The checkpoint's `chat_template.jinja` is byte-identical to the file Qwen ships with Qwen3-8B-Base (SHA-256 `87a2728cb8dc9fe4...`), and its no-tools branch opens a system turn only when the caller supplies a system message; with none, it emits none. The perturbation is that template's default output. Second, the identity of what the model actually conditioned on was checked at the level the server ran, not at the level we constructed: the formatted text and token ids match our offline NO-SYSTEM and EXACT prompts for all 500 problems, the server's own `/tokenize` endpoint agrees for all 500, the `prompt_logprobs` returned with the generations agree, and `usage.prompt_tokens` agrees on all 1,000 generations.

The termination numbers (Table 44) reproduce the offline measurement within sampling noise. Accuracy does not: it falls 3.8 points at the first seed (3.1 over both), a significant drop, where the offline comparison found 0.6 points and no significant difference. The two runs use different decoding seeds and the effect is near the detection threshold at n=500; we report both rather than choosing. The claim the data supports is that the termination effect is an order of magnitude larger than any accuracy change, not that accuracy is untouched.

A13.3 A LARGER BUDGET

The MATH-500 budget (8,192) is below the teacher's median thought length, so a capped generation could be a legitimately long solution. Resampling the same 500 problems at a 30,000-token budget (one seed; not continuations of the 8,192 runs) shows that it is not (Table 45): under EXACT the extra budget lets unfinished attempts finish (solved-then-capped $2.4 \to 1.0\%$), under NO-SYSTEM solved-then-capped generations stay ($19.6 \to 17.4\%$) and more of the capped ones loop. Of the 98 problems solved-then-capped under NO-SYSTEM at 8,192, 54 are still capped at 30,000 (51 of them in a loop).

Table 45: The CE student on MATH-500 at two budgets. NO-SYSTEM rows give the paired difference to EXACT at the same budget. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

budget	format	n	cap	acc	solved-then-capped	loop	well-formed	med. tokens
8,192	EXACT	500	15.6	83.0	2.4	1.0	84.4	1,549
	NO-SYSTEM	500	33.0 ($+17.4^*$)	82.4 (-0.6)	19.6 ($+17.2^*$)	16.6 ($+15.6^*$)	54.0 (-30.4^*)	1,909
30,000	EXACT	500	10.0	85.4	1.0	4.0	90.0	1,484
	NO-SYSTEM	500	26.4 ($+16.4^*$)	84.8 (-0.6)	17.4 ($+16.4^*$)	22.8 ($+18.8^*$)	56.4 (-33.6^*)	1,933

A13.4 SECOND TRAINING SEED

Table 46: The factorial cells of Table 42 for two independently trained CE students (training seeds 42 and 123; same data, recipe and decoding seeds). MATH-500 one decoding seed (n=500), AIME four seeds (n=120). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	training seed	cap	acc	solved-then-capped	loop	well-formed
MATH-500, EXACT	seed 42	15.6	83.0	2.4	1.0	84.4
MATH-500, EXACT	seed 123	16.6	81.0	2.8	1.0	83.4
MATH-500, NO-SYSTEM	seed 42	33.0	82.4	19.6	16.6	54.0
MATH-500, NO-SYSTEM	seed 123	45.6	80.6	29.0	27.8	41.2
MATH-500, ALT-SYSTEM	seed 42	16.0	82.4	3.0	1.4	83.8
MATH-500, ALT-SYSTEM	seed 123	16.8	80.4	2.2	1.0	83.0
MATH-500, NO-SYSTEM+ALT-INSTR	seed 42	31.2	79.4	15.4	16.6	54.4
MATH-500, NO-SYSTEM+ALT-INSTR	seed 123	45.8	72.2	22.8	31.6	34.4
AIME 2025, EXACT	seed 42	52.5	29.2	2.5	10.8	47.5
AIME 2025, EXACT	seed 123	50.8	27.5	0.0	11.7	49.2
AIME 2025, NO-SYSTEM	seed 42	62.5	24.2	8.3	17.5	19.2
AIME 2025, NO-SYSTEM	seed 123	68.3	25.8	12.5	28.3	18.3
AIME 2025, ALT-SYSTEM	seed 42	57.5	23.3	1.7	19.2	42.5

A13.5 BEYOND MATHEMATICS

Table 47: Eight students under EXACT and NO-SYSTEM on GPQA-Diamond (four seeds, budget 24,576), LiveCodeBench problems from January–April 2025 (four seeds, budget 24,576; split into the 112 standard-input problems and the 70 problems with starter code, a split fixed before the coding student's results were read), AIME 2025 and MATH-500 for the two students trained outside mathematics. Each NO-SYSTEM cell gives the paired difference to the same student's EXACT cell. LiveCodeBench has no answer judge here, so its accuracy column is blank. The science set holds 30k physics, chemistry and biology traces; the code set 29,884 competitive-programming traces from a different teacher; both use the recipe and single format of every other student. Repair D is trained from the seed-42 student (its GPQA cells: Table 50). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	bench	n	cap EXACT	cap NO-SYSTEM	loop EXACT	loop NO-SYSTEM	wf EXACT	wf NO-SYSTEM	acc EXACT / NO-SYSTEM
Qwen3-8B CE (seed 42)	GPQA-Diamond	792	27.0	55.1 (+28.0*)	18.1	44.8 (+26.8*)	73.0	24.1 (−48.9*)	20.8 / 23.9
	LCB, stdin	448	43.5	75.2 (+31.7*)	13.6	40.0 (+26.3*)	56.5	13.4 (−43.1*)	–
	LCB, starter code	280	37.1	56.1 (+18.9*)	16.8	31.4 (+14.6*)	62.9	25.4 (−37.5*)	–
Qwen3-8B CE (seed 42) + repair D	GPQA-Diamond	792	29.4	30.2 (+0.8)	20.8	22.6 (+1.8)	70.6	69.7 (−0.9)	17.8 / 17.8
	LCB, stdin	448	47.5	46.0 (−1.6)	16.3	16.7 (+0.4)	52.5	53.8 (+1.3)	–
	LCB, starter code	280	40.4	47.5 (+7.1)	20.4	23.9 (+3.6)	59.6	36.8 (−22.9*)	–
Qwen3-8B CE (seed 123)	GPQA-Diamond	792	26.5	59.7 (+33.2*)	18.4	51.4 (+33.0*)	73.5	13.6 (−59.8*)	19.9 / 25.5
	LCB, stdin	448	46.9	71.4 (+24.6*)	14.7	36.8 (+22.1*)	53.1	11.6 (−41.5*)	–
	LCB, starter code	280	36.4	60.4 (+23.9*)	18.6	35.4 (+16.8*)	63.6	16.8 (−46.8*)	–
Qwen2.5-7B CE	GPQA-Diamond	792	29.7	43.6 (+13.9*)	17.6	29.3 (+11.7*)	70.3	56.2 (−14.1*)	9.0 / 11.2
	LCB, stdin	448	45.3	68.1 (+22.8*)	20.8	41.5 (+20.8*)	54.7	31.2 (−23.4*)	–
	LCB, starter code	280	57.1	69.6 (+12.5*)	35.7	45.7 (+10.0*)	42.9	30.4 (−12.5*)	–
Llama-3.1-8B (base) CE	GPQA-Diamond	792	23.0	27.1 (+4.2*)	13.6	18.6 (+4.9*)	77.0	72.7 (−4.3*)	11.9 / 10.7
	LCB, stdin	448	35.0	42.4 (+7.4*)	17.0	22.1 (+5.1)	64.1	57.6 (−6.5*)	–
	LCB, starter code	280	28.9	40.0 (+11.1*)	12.1	19.6 (+7.5*)	70.7	60.0 (−10.7*)	–
Mistral-7B-v0.3 CE	GPQA-Diamond	792	28.9	28.0 (−0.9)	11.4	12.0 (+0.6)	71.1	71.8 (+0.8)	6.4 / 6.6
	LCB, stdin	448	43.8	44.2 (+0.4)	12.7	13.4 (+0.7)	56.2	55.8 (−0.4)	–
	LCB, starter code	280	45.4	43.9 (−1.4)	15.0	13.2 (−1.8)	54.6	56.1 (+1.4)	–
Qwen3-8B, science-trained	GPQA-Diamond	792	15.4	60.7 (+45.3*)	15.9	51.8 (+35.9*)	71.0	16.2 (−54.8*)	0.0 / 0.5
	LCB, stdin	448	12.9	77.9 (+65.0*)	12.5	68.3 (+55.8*)	71.7	11.2 (−60.5*)	–
	LCB, starter code	280	6.8	81.8 (+75.0*)	7.9	79.6 (+71.8*)	76.8	13.6 (−63.2*)	–
	AIME 2025	240	23.3	81.7 (+58.3*)	15.4	70.4 (+55.0*)	65.4	5.0 (−60.4*)	14.6 / 13.8
	MATH-500	500	4.8	60.2 (+55.4*)	2.4	42.2 (+39.8*)	95.2	29.6 (−65.6*)	49.0 / 58.0
Qwen3-8B, coding-trained	GPQA-Diamond	792	71.7	65.7 (−6.1*)	64.0	59.0 (−5.1*)	18.2	14.6 (−3.5)	1.3 / 3.8
	LCB, stdin	448	55.6	55.8 (+0.2)	25.9	27.7 (+1.8)	44.2	42.4 (−1.8)	–
	LCB, starter code	280	26.8	35.0 (+8.2*)	13.2	21.8 (+8.6*)	72.1	60.0 (−12.1*)	–
	AIME 2025	240	67.1	74.6 (+7.5*)	35.8	57.5 (+21.7*)	24.6	12.9 (−11.7*)	15.4 / 17.5
	MATH-500	500	29.0	33.8 (+4.8*)	1.2	9.2 (+8.0*)	32.6	31.8 (−0.8)	51.6 / 67.0

Executed code. Table 48 runs the last fenced Python block of every LiveCodeBench generation against the public and private tests with the official checker semantics (model code in a resource-limited sandbox). "Solved-then-capped" (r1) counts budget-capped generations whose code already passes. For the brittle Qwen3 students r1 rises from $\leq$0.4 to 5–9% without the system turn

(Qwen2.5-7B and Llama base: $\leq 0.5\%$), almost entirely code written inside the unclosed thought; repair D keeps it at 0.4 vs 0.7%. Pass@1 is not a capability measure here: 35–71% of generations contain no extractable code, and under EXACT the math- and science-trained Qwen3 students often stop standard-input problems without writing any, so their NO-SYSTEM pass@1 is higher on those problems and lower on starter-code problems.

Table 48: LiveCodeBench, executed: pass@1 and solved-then-capped (r1) under EXACT and NO-SYSTEM, paired differences with 95% CIs, overall and by problem format (182 problems, four seeds). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	split	n	pass@1 EXACT	pass@1 NO-SYSTEM	95% CI	r1 EXACT	r1 NO-SYSTEM	95% CI
Qwen3-8B CE (seed 42)	all	728	7.8	13.0 (+5.2*)	[+2.2, +8.5]	0.0	5.2 (+5.2*)	[+3.2, +7.6]
	AtCoder (stdin)	448	3.8	14.3 (+10.5*)	[+6.5, +15.2]	0.0	6.5 (+6.5*)	[+3.6, +9.8]
	LeetCode (function)	280	14.3	11.1 (−3.2)	[-7.1, +0.4]	0.0	3.2 (+3.2*)	[+1.1, +5.7]
Qwen3-8B CE (seed 42) + repair D	all	728	8.5	11.5 (+3.0*)	[+0.7, +5.5]	0.4	0.7 (+0.3)	[-0.4, +1.1]
	AtCoder (stdin)	448	6.5	12.3 (+5.8*)	[+2.7, +9.4]	0.2	0.2 (+0.0)	[-0.7, +0.7]
	LeetCode (function)	280	11.8	10.4 (−1.4)	[-5.0, +2.1]	0.7	1.4 (+0.7)	[-1.1, +2.5]
Qwen3-8B CE (seed 123)	all	728	8.0	11.8 (+3.8*)	[+0.8, +7.0]	0.4	6.0 (+5.6*)	[+3.4, +7.8]
	AtCoder (stdin)	448	4.2	14.1 (+9.8*)	[+5.8, +14.3]	0.0	7.6 (+7.6*)	[+4.7, +10.9]
	LeetCode (function)	280	13.9	8.2 (−5.7*)	[-9.6, -2.1]	1.1	3.6 (+2.5)	[+0.0, +5.4]
Qwen2.5-7B CE	all	728	2.7	2.9 (+0.1)	[-1.1, +1.4]	0.0	0.5 (+0.5*)	[+0.1, +1.1]
	AtCoder (stdin)	448	1.6	1.8 (+0.2)	[-1.3, +1.8]	0.0	0.4 (+0.4)	[+0.0, +1.1]
	LeetCode (function)	280	4.6	4.6 (+0.0)	[-1.8, +1.8]	0.0	0.7 (+0.7)	[+0.0, +1.8]
Llama-3.1-8B (base) CE	all	728	1.8	2.6 (+0.8)	[-0.4, +2.3]	0.0	0.0 (+0.0)	[+0.0, +0.0]
	AtCoder (stdin)	448	0.9	1.6 (+0.7)	[-0.4, +2.0]	0.0	0.0 (+0.0)	[+0.0, +0.0]
	LeetCode (function)	280	3.2	4.3 (+1.1)	[-1.8, +4.3]	0.0	0.0 (+0.0)	[+0.0, +0.0]
Mistral-7B-v0.3 CE	all	728	2.1	1.0 (−1.1)	[-2.3, +0.0]	0.1	0.1 (+0.0)	[-0.4, +0.4]
	AtCoder (stdin)	448	1.6	0.7 (−0.9)	[-2.2, +0.2]	0.2	0.0 (−0.2)	[-0.7, +0.0]
	LeetCode (function)	280	2.9	1.4 (−1.4)	[-3.9, +0.4]	0.0	0.4 (+0.4)	[+0.0, +1.1]
Qwen3-8B, science-trained	all	728	10.7	13.5 (+2.7)	[+0.0, +5.8]	0.0	8.8 (+8.8*)	[+6.2, +11.8]
	AtCoder (stdin)	448	11.4	18.1 (+6.7*)	[+3.3, +10.3]	0.0	11.4 (+11.4*)	[+7.4, +16.1]
	LeetCode (function)	280	9.6	6.1 (−3.6)	[-8.6, +0.7]	0.0	4.6 (+4.6*)	[+2.1, +7.9]
Qwen3-8B, coding-trained	all	728	18.3	17.6 (−0.7)	[-2.6, +1.2]	0.4	0.1 (−0.3)	[-0.8, +0.3]
	AtCoder (stdin)	448	19.4	17.0 (−2.5)	[-5.1, +0.0]	0.2	0.2 (+0.0)	[-0.7, +0.7]
	LeetCode (function)	280	16.4	18.6 (+2.1)	[-0.7, +5.0]	0.7	0.0 (−0.7)	[-1.8, +0.0]

A pre-declared endpoint. To replace the informal notion of brittleness we fix one endpoint: a student is brittle on a benchmark if the paired NO-SYSTEM−EXACT difference in cap exceeds 5 points and is significant (Table 49; full MATH-500 outcome cells for seven students in Table 43). On MATH-500, 7 of the 8 Qwen-derived CE students meet it ($+16.6$ to $+55.4$); the exception is the code-trained student ($+4.8$). No non-Qwen student does ($+0.0$ to $+0.7$). The mean difference between the two groups is $+32.8$ points. Training seeds and data variants of the same base are not independent models: the eight Qwen-derived students are eight training runs over three pretrained initializations (five on Qwen3-8B-Base, two on Qwen2.5-7B, one on DeepSeek-R1-Distill-Qwen-32B), so we take the initialization as the unit of inference. The two groups then separate completely—every Qwen-derived initialization is above every non-Qwen one—and the difference-in-differences is $+35.0$ points on MATH-500 and $+21.9$ on AIME. The exact one-sided permutation test is 3 vs 3 and cannot fall below $1/\binom{6}{3} = 0.05$, which is the value it attains on both; on GPQA and LiveCodeBench only two initializations per group have cells, so its floor there is 0.167. The family pattern is an association over six initializations, not an established effect of pretraining lineage.

Table 49: Δcap (NO-SYSTEM$-$EXACT, points) with 95% CIs. Bold: meets the endpoint; ℓ: also meets it for the loop rate. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student (training run)	MATH-500	AIME 2025	GPQA-D	LCB (all)	LCB AtCoder	LCB LeetCode
Qwen-derived						
initialization: Qwen3-8B-Base – 5 CE training runs of one pretrained model, not 5 models, plus 1 repair arm (not in the DiD)						
Qwen3-8B (base), seed 42	$+16.6^{\ell}$ [+13.2, +20.0]	$+17.5^{\ell}$ [+9.2, +26.4]	$+28.0^{\ell}$ [+23.1, +33.2]	$+26.8^{\ell}$ [+21.8, +31.7]	$+31.7^{\ell}$ [+25.7, +37.9]	$+18.9^{\ell}$ [+11.1, +26.8]
Qwen3-8B (base), seed 123	$+29.0^{\ell}$ [+24.0, +33.8]	$+17.5^{\ell}$ [+5.0, +30.0]	$+33.2^{\ell}$ [+28.5, +37.9]	$+24.3^{\ell}$ [+19.4, +29.4]	$+24.6^{\ell}$ [+17.6, +31.2]	$+23.9^{\ell}$ [+16.1, +31.1]
Qwen3-8B (base), 1 epoch	$+48.5^{\ell}$ [+44.7, +52.3]	$+36.7^{\ell}$ [+26.7, +47.1]	–	–	–	–
Qwen3-8B (base), science-trained	$+55.4^{\ell}$ [+50.8, +60.0]	$+58.3^{\ell}$ [+48.3, +68.3]	$+45.3^{\ell}$ [+41.0, +49.6]	$+68.8^{\ell}$ [+64.6, +72.9]	$+65.0^{\ell}$ [+59.4, +70.3]	$+75.0^{\ell}$ [+68.9, +81.4]
Qwen3-8B (base), coding-trained	$+4.8^{\ell}$ [+0.6, +9.0]	$+7.5^{\ell}$ [+1.2, +13.3]	-6.1 [−10.4, −1.8]	$+3.3$ [−0.7, +7.4]	$+0.2$ [−4.7, +4.9]	$+8.2^{\ell}$ [+1.4, +15.7]
Qwen3-8B seed 42 + repair D	$+1.6$ [−0.3, +3.5]	$+3.3$ [−0.6, +7.5]	$+0.8$ [−3.0, +4.5]	$+1.8$ [−2.3, +6.0]	-1.6 [−6.5, +3.1]	$+7.1$ [+0.0, +14.3]
initialization: Qwen2.5-7B – 2 CE training runs of one pretrained model, not 2 models						
Qwen2.5-7B (base), seed 42	$+40.0^{\ell}$ [+36.3, +43.7]	$+7.1^{\ell}$ [−1.7, +16.7]	$+13.9^{\ell}$ [+9.5, +18.4]	$+18.8^{\ell}$ [+14.1, +23.4]	$+22.8^{\ell}$ [+16.7, +29.0]	$+12.5^{\ell}$ [+5.7, +19.3]
Qwen2.5-7B (base), seed 123	$+31.0^{\ell}$ [+27.3, +34.8]	$+9.2$ [+1.2, +17.1]	–	–	–	–
DS-R1-Distill-Qwen-32B	$+39.6^{\ell}$ [+35.2, +44.0]	$+30.8^{\ell}$ [+17.5, +45.0]	–	–	–	–
non-Qwen						
Llama-3.1-8B (base)	$+0.0$ [−3.0, +3.0]	$+2.5$ [−3.8, +8.8]	$+4.2$ [+0.3, +8.0]	$+8.8^{\ell}$ [+4.7, +12.8]	$+7.4$ [+2.0, +12.9]	$+11.1^{\ell}$ [+5.4, +16.8]
Llama-3.1-8B-Instruct	$+0.2$ [−3.4, +3.8]	-2.5 [−10.8, +5.8]	–	–	–	–
Mistral-7B-v0.3 (base)	$+0.7$ [−2.1, +3.5]	$+0.8$ [−6.2, +7.9]	-0.9 [−4.7, +2.9]	-0.3 [−4.0, +3.6]	$+0.4$ [−4.5, +5.4]	-1.4 [−7.5, +4.6]
Difference-in-differences of Δcap (Qwen-derived CE − non-Qwen CE); primary statistic						
DiD [95% CI]	$+35.0$ [+32.4, +37.6]	$+21.9$ [+15.0, +28.6]	$+17.9$ [+14.2, +21.7]	$+20.6$ [+16.8, +24.4]	$+22.7$ [+17.8, +27.7]	$+17.2$ [+11.5, +22.7]
initializations n/n; one-sided perm. p (floor)	3/3; 0.050 (0.050)	3/3; 0.050 (0.050)	2/2; 0.167 (0.167)	2/2; 0.167 (0.167)	2/2; 0.167 (0.167)	2/2; 0.167 (0.167)
runs $n_{\text{Qwen}}/n_{\text{non}}$; one-sided perm. p	8/3; 0.006	8/3; 0.006	5/2; 0.143	5/2; 0.095	5/2; 0.143	5/2; 0.095

A13.6 GPQA-DIAMOND

Table 50: The math-trained CE student and its arm-D repair on GPQA-Diamond (198 four-option science questions, budget 24,576, four decoding seeds; accuracy = last boxed letter, near the 25% chance level under every format). Paired differences to the CE student as in Table 77. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	model	n	cap	acc	solved-then-capped	loop	well-formed
EXACT	CE student	792	27.0	20.8	0.3	18.1	73.0
	D format-augmented replay	792	29.4 (+2.4)	17.8 (−3.0*)	0.5 (+0.3)	20.8 (+2.8)	70.6 (−2.4)
NO-SYSTEM	CE student	792	55.1	23.9	11.5	44.8	24.1
	D format-augmented replay	792	30.2 (−24.9*)	17.8 (−6.1*)	0.1 (−11.4*)	22.6 (−22.2*)	69.7 (+45.6*)
HISTORICAL	CE student	792	58.7	23.6	11.1	49.6	24.4

A14 MECHANISM: WHERE THE CHANGE SITS AND WHAT REMOVES IT

A14.1 POSITION-WISE KL PROBE

Over the thought body no student's predictions move by more than 0.008 nats under NO-SYSTEM; at a closing token the Qwen3-8B and Qwen2.5-7B CE students move by 1.4 to 10.9 (Table 51).

Where the 32B student diverges. On its own natural-stop trajectories the 32B student's KL at `</think>` is negligible (median 2.3×10^{-5}, against 1.69 for Qwen3-8B, on item-matched EXACT trajectories), while its per-token KL over the reasoning (excluding the closing and end tokens) is about twice the 8B's in deciles 2–8 of those trajectories; on its own NO-SYSTEM trajectories it rises again over the last two deciles, which is mostly the answer segment (answer KL 0.110 vs 0.043 in

Table 51: Where the perturbed format changes the student's predictions. Teacher-forced on 108–189 held-out teacher traces (182 for the Qwen students; median 9k tokens): mean per-token $\mathrm{KL}(p_{\mathrm{EXACT}} \,\|\, p_{\mathrm{perturbed}})$ in nats over the thought body and the answer segment, at the </think> token and at the end-of-turn token, and the median probability of the end-of-turn token at the teacher's terminal state under the two formats. Repair arms start from the seed-42 CE student. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	perturbed format	traces	think body	answer	</think>	end-of-turn	p(end-of-turn), median
CE student (seed 42)	NO-SYSTEM	182	0.005	0.002	1.41	0.71	0.968 → 0.638
CE student (seed 123)	NO-SYSTEM	182	0.006	0.002	2.20	0.95	0.968 → 0.541
CE student (seed 42)	ALT-SYSTEM	182	0.003	0.002	0.13	0.12	0.968 → 0.916
sharpened student	NO-SYSTEM	182	0.005	0.002	2.82	1.81	0.988 → 0.350
Llama-3.1-8B-Instruct CE student	NO-SYSTEM	189	0.002	0.001	0.00	0.02	0.977 → 0.988
Llama-3.1-8B (base) CE student	NO-SYSTEM	189	0.003	0.002	0.00	0.09	0.973 → 0.964
Qwen2.5-7B (base) CE student	NO-SYSTEM	182	0.001	0.002	0.00	9.62	0.979 → 0.000
Qwen2.5-7B (base) CE student, seed 123	NO-SYSTEM	182	0.001	0.002	0.00	10.90	0.976 → 0.000
CE student, 1 ep × 30k	NO-SYSTEM	182	0.006	0.003	2.68	2.69	0.185 → 0.004
Mistral-7B-v0.3 (base) CE student	NO-SYSTEM	175	0.001	0.001	0.00	0.00	0.982 → 0.984
DS-R1-Distill-Qwen-32B CE student	NO-SYSTEM	108	0.008	0.005	0.00	0.01	0.971 → 0.970
repair B (teacher-prefix terminal)	NO-SYSTEM	182	0.003	0.002	0.03	0.27	0.949 → 0.931
repair C (replay only)	NO-SYSTEM	182	0.005	0.002	1.34	0.66	0.969 → 0.662
repair D (format-augmented replay)	NO-SYSTEM	182	0.002	0.001	0.04	0.11	0.952 → 0.896

Table 52: The same probe on base checkpoints before any SFT (teacher traces, as above) and on the students' own trajectories (150 natural stops under each format). The last column is the median probability of </think> at the closing position under EXACT and under NO-SYSTEM. **Read this column only within a row.** It is recorded at the *last* sub-token of the spelling, which for tokenizers that spell </think> in several pieces is the probability of finishing a token already begun, not the decision to begin closing. Measured at the first sub-token instead, Mistral-7B-v0.3 reads 4.5×10^{-5} rather than 0.976, and the apparent separation between brittle and robust bases in this column disappears. §4 reports the decision-point quantity. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

checkpoint	histories	traces	think body	answer	</think>	end-of-turn	p(</think>), median
Qwen3-8B-Base (no SFT)	teacher traces	182	0.005	0.005	0.97	0.09	0.000 → 0.000
Qwen2.5-7B (no SFT)	teacher traces	182	0.018	0.008	0.04	0.06	0.141 → 0.182
Qwen2.5-Math-7B (no SFT)	teacher traces	182	0.033	0.099	0.11	0.12	0.000 → 0.000
DS-R1-Distill-Qwen-32B (no local SFT)	teacher traces	108	0.005	0.003	0.01	0.00	0.998 → 0.991
Llama-3.1-8B (no SFT)	teacher traces	189	0.007	0.003	0.11	0.03	0.598 → 0.425
Llama-3.1-8B-Instruct (no SFT)	teacher traces	189	0.012	0.005	0.05	0.00	0.572 → 0.522
Mistral-7B-v0.3 (no SFT)	teacher traces	175	0.002	0.001	0.01	0.00	0.976 → 0.980
Qwen3-8B CE student	own EXACT stops	150	0.020	0.002	2.47	0.27	0.995 → 0.113
Qwen3-8B CE student	own NO-SYSTEM stops	150	0.026	0.003	1.05	0.19	0.994 → 0.411
DS-R1-Distill-Qwen-32B CE student	own EXACT stops	150	0.019	0.008	0.00	0.00	1.000 → 1.000
DS-R1-Distill-Qwen-32B CE student	own NO-SYSTEM stops	150	0.043	0.110	0.00	0.05	1.000 → 1.000
Qwen3-8B science-trained student	own EXACT GPQA stops	150	0.015	0.016	1.85	1.64	0.246 → 0.014
Qwen3-8B code-trained student	own EXACT code stops	150	0.004	0.001	0.29	0.04	0.984 → 0.785

the thought; Table 52), and its rate of writing any box falls from 94.8 to 80.6% (Table 53). The NO-SYSTEM format thus changes how the 32B reasons and answers more than when it closes.

We re-ran this probe on *item-matched* trajectories, so the two students are compared on the same problems. Restricting to problems on which a student stops naturally under both formats gives 511 matched pairs for the 8B and 246 for the 32B, and the dissociation survives matching. For the 8B the median KL at </think> is 1.69 nats and the median probability of closing at that position falls from 0.995 under the training format to 0.179 without the system turn. For the 32B the same quantities are 2.3×10^{-5} nats and 0.99993 against 0.99988: on identical problems, the conditional that ends the 8B's thought collapses and the 32B's does not move at all. The 32B probe was run independently on two clusters and the two runs agree to the reported precision. Matching *attenuates* the 8B's collapse (0.113 on 150 unmatched trajectories against 0.179 here) because it conditions on stopping under NO-SYSTEM, a post-treatment outcome, and so drops the most brittle items: the 57 trajectories excluded by matching fall to 0.033, while on the 93 common to both sets the two estimates agree (0.177 against 0.179). The matched figure is therefore the conservative one.

Answer anatomy. Accuracy and the chance of ever writing a correct box barely move for the 8B student (Table 53; 83.6 vs 83.8%), and overwriting a correct box with a wrong final one is rare ($\leq 1.4\%$). What changes is what follows the answer: under NO-SYSTEM, 103 of the 165 capped MATH-500 generations already contain a correct box and spend a median 7,107 further tokens (1,662 under EXACT, 12 of 78 capped rows); per request, 34.6% of all NO-SYSTEM tokens fall after the answer, against 2.8% (Table 54). The judge (math_verify 0.9.0, zero unresolved verdicts) accepts 45–64 equivalent forms per MATH-500 cell that string matching rejects (50 vs 49 for the 8B under the two formats), so it does not create the gap.

Table 53: Answer anatomy: rows with any / an ever-correct / a correct first / a correct last box (last correct = accuracy), overwrites, capped rows and those among them with a correct box, median tokens after the first correct box among capped rows, and last-box verdicts on which math_verify and string matching disagree. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

	any box	ever corr.	first corr.	last corr.	overwrite	capped / ever corr.	tok. after 1st corr.	judge diff.
Qwen3-8B, MATH-500 (n=500), EXACT	88.8	83.6	77.2	83.0	0.6	78 / 12	1,662	50
Qwen3-8B, MATH-500 (n=500), NO-SYSTEM	87.8	83.8	79.4	82.4	1.4	165 / 103	7,107	49
Qwen3-8B, AIME 2025 (n=360), EXACT	55.8	26.1	24.4	26.1	0.0	179 / 4	16,564	0
Qwen3-8B, AIME 2025 (n=360), NO-SYSTEM	54.2	25.0	22.2	24.4	0.6	242 / 33	28,149	1
DS-R1-Distill-Qwen-32B, MATH-500 (n=500), EXACT	94.8	92.8	86.2	92.2	0.6	50 / 24	2,067	64
DS-R1-Distill-Qwen-32B, MATH-500 (n=500), NO-SYSTEM	80.6	79.8	74.4	79.4	0.4	248 / 150	5,464	45

Table 54: Cost per request, MATH-500, first decoding seed. *mean* and *total* are tokens emitted per generation and over the 500-problem run; *after answer* counts tokens emitted after the first correct `\boxed{}`, over the generations that reach the budget with a correct answer already written. The mean ratio between the two formats is modest (1.35× for the seed-42 student); what changes is where the tokens go—34.6% of all tokens emitted under NO-SYSTEM fall after the answer, against 2.8% under the training format, and the post-answer increase (+621,257) exceeds the increase in total tokens (+491,785). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

cell	n	cap	med. tok	mean tok	total tok	mean/EXACT	solved-then-capped	mean tok after	sum tok after
MATH-500, CE student, seed 42, EXACT	500	15.6	1,549	2,837	1,418,522	–	12	3,315	39,782
MATH-500, CE student, seed 42, NO-SYSTEM	500	33.0	1,909	3,821	1,910,307	1.35	103	6,418	661,039
MATH-500, CE student, seed 42, HISTORICAL	500	24.8	1,202	3,044	1,521,854	1.07	67	6,758	452,754
MATH-500, CE student, seed 42 (cluster stack, 2 seeds), EXACT	1000	15.7	1,549	2,834	2,834,409	–	25	3,638	90,946
MATH-500, CE student, seed 42 (cluster stack, 2 seeds), NO-SYSTEM	1000	30.9	1,938	3,699	3,699,208	1.31	176	6,537	1,150,489
MATH-500, CE student, seed 123, EXACT	500	16.6	1,499	2,835	1,417,413	–	14	4,544	63,613
MATH-500, CE student, seed 123, NO-SYSTEM	500	45.6	4,379	4,624	2,312,182	1.63	153	7,054	1,079,268
MATH-500, repair D (format-augmented replay), EXACT	1000	14.8	1,465	2,733	2,733,381	–	27	2,858	77,159
MATH-500, repair D (format-augmented replay), NO-SYSTEM	1000	16.4	1,559	2,894	2,893,609	1.06	25	3,590	89,758
AIME 2025, CE student, seed 42 (cluster stack, 12 seeds), EXACT	360	49.7	29,638	19,973	7,190,452	–	4	15,284	61,136
AIME 2025, CE student, seed 42 (cluster stack, 12 seeds), NO-SYSTEM	360	67.2	30,000	23,696	8,530,582	1.19	33	23,362	770,936
AIME 2025, CE student, seed 42 (cluster stack, 12 seeds), HISTORICAL	360	58.3	30,000	21,790	7,844,392	1.09	26	26,974	701,315
AIME 2025, CE student, seed 123, EXACT	120	50.8	30,000	20,216	2,425,974	–	0	–	0
AIME 2025, CE student, seed 123, NO-SYSTEM	120	68.3	30,000	23,845	2,861,454	1.18	17	26,542	451,210
AIME 2025, repair D (format-augmented replay), EXACT	360	47.5	25,557	19,423	6,992,449	–	0	–	0
AIME 2025, repair D (format-augmented replay), NO-SYSTEM	360	50.8	30,000	20,269	7,296,676	1.04	3	19,849	59,546

A14.2 STOP HAZARD BY POSITION

Figure 10 plots the per-bin probability (bins are unequal: MATH-500 edges at 0, 0.5k, 1k, 1.5k, 2k, 3k, 4k, 6k, 8,192 tokens; AIME at 0, 4k, 8k, 12k, 16k, 20k, 25k, 30k) of a natural stop among generations still running at the start of the bin (budget-capped generations right-censored). HISTORICAL shows the same decay as NO-SYSTEM, a different well-formed system prompt does not, the 32B student's hazard rises with length under its training format and decays without the system turn, and the Llama student's two curves coincide. These are population hazards—later bins contain only the generations still running—so they describe the aggregate rather than an individual trajectory.

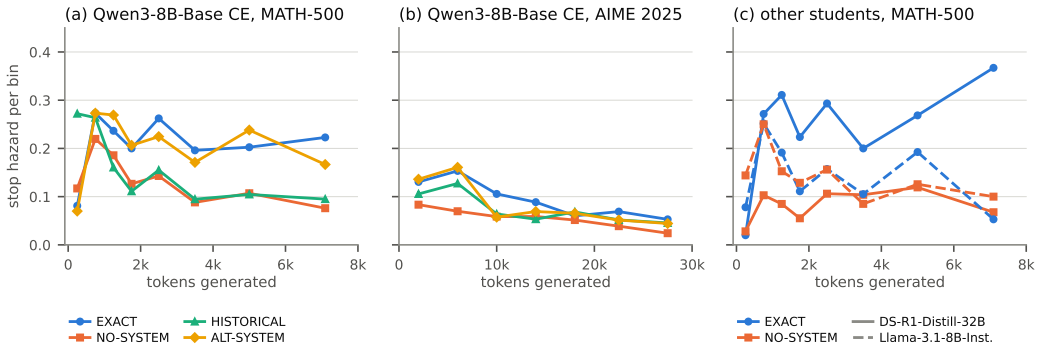

Figure 10: Stop hazard by position: the fraction of generations still running at the start of a token bin that stop naturally within it (budget-capped generations are right-censored). Under the training format (and under a different but well-formed system prompt) the hazard of the 8B student is roughly flat after the first 500 tokens; without the system turn it decays with length, from 0.22 to 0.08 per bin on MATH-500 (two decoding seeds; AIME twelve seeds). The 32B student shows the same signature more strongly; the Llama student's hazard is the same with and without the system turn.

A14.3 PREFIX-SWAP RESCUE

Table 55: 96 budget-capped NO-SYSTEM MATH-500 generations of the 8B CE student, drawn at random (fixed seed) from all of its capped ones, truncated at the given share of their length and continued (budget 8,192 total, one seed) under the same NO-SYSTEM prefix (control, a re-sample of the model's own continuation) or under the EXACT prefix (swap). Parentheses: paired difference swap−control. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prefix kept	n	continue under NO-SYSTEM: cap	solved-then-capped	loop	wf	continue under EXACT: cap	solved-then-capped	loop	wf
10%	96	84.4	41.7	34.4	11.5	53.1 (−31.2*)	15.6 (−26.0*)	14.6 (−19.8*)	46.9 (+35.4*)
25%	96	83.3	40.6	39.6	9.4	75.0 (−8.3)	35.4 (−5.2)	32.3 (−7.3*)	24.0 (+14.6*)
50%	96	95.8	53.1	43.8	3.1	92.7 (−3.1)	47.9 (−5.2)	43.8	6.2 (+3.1)
75%	96	94.8	49.0	44.8	3.1	94.8	49.0	44.8	4.2 (+1.0)

A14.4 CLOSURE-LOGIT SPLICE: PRE-REGISTERED RUN AND THE 32B STUDENT

Table 56: Causal test of the localization: free-running generation (all 500 MATH-500 problems, budget 8,192, one seed, one sampler for every row) in which at every step the probabilities of the named tokens are replaced by those the same model assigns under the other format given the same history, all other tokens rescaled. Malformed stop: turn ended with the thought unclosed. Parentheses: paired difference to plain NO-SYSTEM sampling. Paired against plain EXACT, the forward splices differ on no metric but +0.8 points malformed stops for `</think>` alone; the reverse splice by −41.6/+42.4/−0.8 points well-formed/malformed/cap (*). First run: Table 60. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	500	31.6	80.8	15.4	14.2	53.8	14.6
EXACT (plain sampling)	500	16.6 (−15.0*)	82.8 (+2.0)	3.0 (−12.4*)	1.0 (−13.2*)	83.4 (+29.6*)	0.0 (−14.6*)
NO-SYSTEM + `</think>` and end-of-turn	500	16.6 (−15.0*)	82.4 (+1.6*)	3.0 (−12.4*)	1.2 (−13.0*)	83.4 (+29.6*)	0.0 (−14.6*)
NO-SYSTEM + sham (`Wait`, break)	500	30.0 (−1.6)	80.8	15.0 (−0.4)	14.4 (+0.2)	55.0 (+1.2)	15.0 (+0.4)
NO-SYSTEM + sham 2 (`the, is`)	500	32.8 (+1.2)	81.8 (+1.0)	18.4 (+3.0)	17.0 (+2.8)	53.0 (−0.8)	14.2 (−0.4)
NO-SYSTEM + `</think>` only	500	17.2 (−14.4*)	81.2 (+0.4)	3.0 (−12.4*)	1.2 (−13.0*)	82.0 (+28.2*)	0.8 (−13.8*)
EXACT + `</think>` from NO-SYSTEM (reverse)	500	15.8 (−15.8*)	82.8 (+2.0)	2.4 (−13.0*)	0.6 (−13.6*)	41.8 (−12.0*)	42.4 (+27.8*)

Table 57 gives the gated and constant-bias arms on the Qwen3 CE student and Table 58 the end-of-turn arms on the Qwen2.5 student (first 200 problems of the same order, one seed). The gate opens where the temperature-scaled $p_{\text{EXACT}}(\texttt{</think>})$ exceeds τ; it fired on 0.83 steps per generation on average (0.08% of steps) at either threshold. Table 59 measures how far the spliced tokens move

between the formats over the thought body: the sham tokens shift more than the closing tokens do there, so their null effect is not for lack of a difference to splice.

Table 57: Gated </think> splice and constant logit bias, Qwen3 CE student, MATH-500 (200 problems). Paired differences to plain NO-SYSTEM sampling. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	200	36.0	76.0	15.0	13.0	49.5	14.5
EXACT (plain sampling)	200	19.5 (−16.5*)	77.5 (+1.5)	1.5 (−13.5*)	1.5 (−11.5*)	80.5 (+31.0*)	0.0 (−14.5*)
NO-SYSTEM + gated </think> splice, τ=0.05	200	19.5 (−16.5*)	77.5 (+1.5)	2.0 (−13.0*)	2.0 (−11.0*)	80.5 (+31.0*)	0.0 (−14.5*)
NO-SYSTEM + gated </think> splice, τ=0.2	200	19.5 (−16.5*)	77.5 (+1.5)	2.0 (−13.0*)	2.0 (−11.0*)	80.5 (+31.0*)	0.0 (−14.5*)
NO-SYSTEM + constant </think> logit bias +1	200	30.0 (−6.0*)	77.5 (+1.5)	10.0 (−5.0*)	8.0 (−5.0*)	64.5 (+15.0*)	5.5 (−9.0*)
NO-SYSTEM + constant </think> logit bias +2	200	28.0 (−8.0*)	78.0 (+2.0*)	8.5 (−6.5*)	5.5 (−7.5*)	71.0 (+21.5*)	1.0 (−13.5*)
NO-SYSTEM + constant </think> logit bias +3	200	28.5 (−7.5*)	78.5 (+2.5*)	9.0 (−6.0*)	6.5 (−6.5*)	71.5 (+22.0*)	0.0 (−14.5*)

Table 58: End-of-turn splices, Qwen2.5-7B CE student, MATH-500 (200 problems). The </think> path splices its first two sub-word pieces. Paired differences to plain NO-SYSTEM sampling. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	200	63.5	71.5	37.0	25.0	36.5	0.0
EXACT (plain sampling)	200	26.5 (−37.0*)	73.0 (+1.5)	5.0 (−32.0*)	1.5 (−23.5*)	73.5 (+37.0*)	0.0
NO-SYSTEM + end-of-turn only	200	28.5 (−35.0*)	71.5	5.0 (−32.0*)	3.5 (−21.5*)	71.5 (+35.0*)	0.0
NO-SYSTEM + </think> path only	200	65.0 (+1.5)	71.0 (−0.5)	39.5 (+2.5)	28.0 (+3.0)	34.5 (−2.0)	0.5 (+0.5)
NO-SYSTEM + sham (Wait, break)	200	64.0 (+0.5)	70.0 (−1.5)	36.5 (−0.5)	26.5 (+1.5)	35.5 (−1.0)	0.5 (+0.5)
EXACT + end-of-turn from NO-SYSTEM (reverse)	200	89.0 (+25.5*)	69.0 (−2.5)	59.5 (+22.5*)	42.5 (+17.5*)	11.0 (−25.5*)	0.0

Table 59: How far each spliced token moves between the formats (teacher-forced on 200 held-out traces, T=0.6, thought body): mean $|p_{\text{EXACT}}(t) - p_{\text{NO-SYSTEM}}(t)|$ over every position, and $|\Delta \log p(t)|$ where t is the realized token, with 95% CIs. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	token t	realized n	$\lvert\Delta p(t)\rvert$, every position	$\lvert\Delta \log p(t)\rvert$ where t realized
Qwen3-8B	</think>	0	4.9e-07 [1.9e-07, 8.7e-07]	–
	end-of-turn	0	1.1e-06 [5.7e-07, 1.8e-06]	–
	Wait	1089	5.5e-05 [5.1e-05, 5.9e-05]	0.107 [0.099, 0.114]
	break	489	1.7e-05 [1.4e-05, 2.0e-05]	0.104 [0.083, 0.127]
	the	14950	0.0008 [0.0007, 0.0008]	0.095 [0.089, 0.100]
	is	13087	0.0005 [0.0005, 0.0005]	0.067 [0.064, 0.070]
	KL($p_{\text{EXACT}} \,\|\, p_{\text{NO-SYSTEM}}$) per position ($n = 1545576$)		0.0063 [0.0061, 0.0066]	–
Qwen2.5-7B	</ (1st piece of </think>)	0	1.0e-06 [3.4e-07, 2.2e-06]	–
	end-of-turn	0	1.0e-10 [1.4e-11, 2.3e-10]	–
	Wait	1089	3.2e-05 [2.9e-05, 3.4e-05]	0.068 [0.063, 0.073]
	break	489	1.0e-05 [8.3e-06, 1.2e-05]	0.061 [0.050, 0.073]
	the	14950	0.0005 [0.0004, 0.0005]	0.058 [0.055, 0.061]
	is	13087	0.0003 [0.0003, 0.0003]	0.042 [0.040, 0.044]
	KL($p_{\text{EXACT}} \,\|\, p_{\text{NO-SYSTEM}}$) per position ($n = 1545758$)		0.0017 [0.0016, 0.0017]	–

Table 60: The pre-registered first splice run (120 problems, one seed, single-row HF sampler; the main-text table reports the batched 500-problem replication). The splice is applied to the temperature-scaled (T=0.6) distributions, before top-p truncation. The sham splices control for the procedure itself—overwriting two tokens' probabilities at every step and renormalizing—not for a perturbation of equal size: Table 59 shows the sham tokens move *more* between the formats over the thought body than the closing tokens do, so their null is not for lack of a difference to splice. Same conventions as Table 56; the last column is the share of natural stops that never closed the thought. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	120	34.2	76.7	14.2	15.8	50.8	15.0
EXACT (plain sampling)	120	20.0 (−14.2*)	80.8 (+4.2)	5.0 (−9.2*)	1.7 (−14.2*)	80.0 (+29.2*)	0.0 (−15.0*)
NO-SYSTEM + </think> and end-of-turn	120	22.5 (−11.7*)	77.5 (+0.8)	4.2 (−10.0*)	5.0 (−10.8*)	77.5 (+26.7*)	0.0 (−15.0*)
NO-SYSTEM + sham (Wait, break)	120	33.3 (−0.8)	76.7	13.3 (−0.8)	15.0 (−0.8)	52.5 (+1.7)	14.2 (−0.8)
NO-SYSTEM + </think> only	120	22.5 (−11.7*)	76.7	4.2 (−10.0*)	5.0 (−10.8*)	75.8 (+25.0*)	1.7 (−13.3*)
NO-SYSTEM + end-of-turn only	120	30.8 (−3.3)	76.7	11.7 (−2.5)	12.5 (−3.3)	57.5 (+6.7*)	11.7 (−3.3)

Table 61: The same splice on the DeepSeek-R1-Distill-Qwen-32B CE student (120 problems, one seed, batch 8, model sharded over two GPUs). Replacing the closing-token probabilities with their EXACT values changes nothing for this student. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	120	54.2	74.2	29.2	19.2	45.8	0.0
EXACT (plain sampling)	120	12.5 (−41.7*)	86.7 (+12.5*)	1.7 (−27.5*)	0.8 (−18.3*)	87.5 (+41.7*)	0.0
NO-SYSTEM + closure splice	120	54.2	74.2	29.2	19.2	45.8	0.0
NO-SYSTEM + sham splice	120	49.2 (−5.0)	75.0 (+0.8)	25.8 (−3.3)	17.5 (−1.7)	50.8 (+5.0)	0.0

A14.5 THE SPLICE ON A SECOND TRAINING SEED AND OUTSIDE MATHEMATICS

Same protocol as Table 56 (MATH-500, the deterministic 200-item subset, budget 8,192, one decoding seed, T=0.6/top-p 0.95): an independently trained Qwen3-8B student (seed 123) and the science-trained student. The intervention fires on 1.7 and 3.1 steps per generation respectively, and moves 6×10^{-4} and 9×10^{-2} of probability mass per step; on the 32B student the same arms move nothing (≤ 0.1 changed steps per generation, mass $< 10^{-5}$), which is why its splice (Table 61) is a null intervention rather than a null effect.

Table 62: Closure splice on the Qwen3-8B CE student trained with seed 123. Paired differences to plain NO-SYSTEM sampling, as in Table 56. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	200	41.0	78.5	21.5	21.5	40.0	19.0
EXACT (plain sampling)	200	17.0 (−24.0*)	80.0 (+1.5)	2.0 (−19.5*)	1.5 (−20.0*)	83.0 (+43.0*)	0.0 (−19.0*)
NO-SYSTEM + </think> and end-of-turn	200	21.5 (−19.5*)	78.5	4.5 (−17.0*)	0.0 (−21.5*)	78.5 (+38.5*)	0.0 (−19.0*)
NO-SYSTEM + sham (Wait, break)	200	42.0 (+1.0)	78.0 (−0.5)	25.0 (+3.5)	24.5 (+3.0)	39.5 (−0.5)	18.5 (−0.5)
EXACT + </think> from NO-SYSTEM (reverse)	200	20.0 (−21.0*)	77.0 (−1.5)	4.5 (−17.0*)	3.0 (−18.5*)	16.5 (−23.5*)	63.5 (+44.5*)

Table 63: Closure splice on the science-trained student (30k physics/chemistry/biology traces). Paired differences to plain NO-SYSTEM sampling. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

decoding condition	n	cap	acc	solved-then-capped	loop	well-formed	malformed
NO-SYSTEM (plain sampling)	200	65.0	45.0	33.0	47.5	14.0	21.0
EXACT (plain sampling)	200	5.0 (−60.0*)	37.0 (−8.0)	0.0 (−33.0*)	3.0 (−44.5*)	61.5 (+47.5*)	33.5 (+12.5*)
NO-SYSTEM + </think> and end-of-turn	200	8.0 (−57.0*)	36.0 (−9.0)	1.5 (−31.5*)	4.5 (−43.0*)	55.5 (+41.5*)	36.5 (+15.5*)
NO-SYSTEM + sham (Wait, break)	200	65.5 (+0.5)	50.0 (+5.0)	35.0 (+2.0)	44.0 (−3.5)	13.5 (−0.5)	21.0
EXACT + </think> from NO-SYSTEM (reverse)	200	5.0 (−60.0*)	24.0 (−21.0*)	0.0 (−33.0*)	3.5 (−44.0*)	24.0 (+10.0*)	71.0 (+50.0*)

A14.6 WHY THE STOP DECISION? A HYPOTHESIS, AND RELATED WORK ON DEGENERATION

In a 10k-token target each closing token is supervised once, against thousands of reasoning tokens seen in many contexts, so the closing decisions have the least coverage of the prefix distribution and most readily fall back on the base prior—to continue. The student's own continuation then compounds the shift (exposure bias; Ranzato et al., 2016; Bengio et al., 2015; not the self-correcting kind, He et al., 2021) into the repetition attractor of autoregressive decoding (Holtzman et al., 2020; Fu et al., 2021; Xu et al., 2022).

Imitation and degeneration. Welleck et al. (2020b) train against repetition directly. Long-trace SFT is sequence-level distillation (Kim & Rush, 2016); our on-policy arm follows dataset aggregation (Ross et al., 2011) and repaired but left stops unclosed, where a pure coverage fix did not.

A15 BOUNDARIES: WHAT CHANGES THE SIZE OF THE EFFECT

A15.1 OPTIMIZER

On AIME the $\beta_1{=}\beta_2$ reduction is as large for two of the three objectives and, for the CE pair, confined to solved-then-capped generations and well-formed stops; well-formed stops on MATH-500 rise by 19–42 points. A hypothesis: with $\beta_1{=}\beta_2$ the bias-corrected adaptive step of every coordinate is bounded by the learning rate ($|m_t| \leq \sqrt{v_t}$, excluding weight decay), which removes the amplification standard Adam gives to parameters with rare, bursty gradients; whether the parameters serving the closing tokens are such parameters we have not tested. The sharpened student under HISTORICAL on AIME: cap 87.5 vs 58.3%, loops 61 vs 20% (four seeds, $n{=}120$, cluster stack); on MATH-500 it caps on 52.0 against 24.8% (Table 65).

Table 64: Matched training seeds, AdamW $(\beta_1, \beta_2) = (0.9, 0.95) \rightarrow (0.95, 0.95)$; each cell shows the standard-optimizer value and, after the arrow, the $\beta_1{=}\beta_2$ value with the paired difference. "CE′" is a confidence-weighted arm (a variant of the sharpened objective's bonus with a self-calibrated scale) whose scale collapsed to ≈ 0, i.e. effectively CE but a separate training run; its EXACT cells are close to the main CE student's (Table 42) but its NO-SYSTEM cap is 12–16 points higher. The CE row uses the main CE checkpoint's local-stack generations, not the cluster-stack cells of Table 42; "sharpened" is the objective of §4. A further arm (the BalancedAdamW series at ratio $\approx$66 in Table 11) that up-weights the stop tokens themselves during SFT flips the off-manifold failure into premature stopping (MATH-500 accuracy $83 \rightarrow$ 62–71% with cap $\leq 14\%$), consistent with the stop decision being the fragile point in either direction. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

objective	bench	format	n	cap	acc	solved-then-capped	loop	well-formed
CE	MATH-500	EXACT	1000	15.0 → 16.5 (+1.5)	83.4 → 81.5 (−1.9)	2.3 → 2.7 (+0.4)	1.3 → 1.4 (+0.1)	85.0 → 83.5 (−1.5)
CE	MATH-500	NO-SYSTEM	1000	31.6 → 22.2 (−9.4*)	82.8 → 81.9 (−0.9)	18.4 → 8.4 (−10.0*)	16.3 → 7.3 (−9.0*)	54.4 → 73.8 (+19.4*)
CE	AIME	EXACT	360	47.5 → 49.4 (+1.9)	26.7 → 26.7	1.1 → 0.3 (−0.8)	9.7 → 11.1 (+1.4)	52.5 → 50.6 (−1.9)
CE	AIME	NO-SYSTEM	120	62.5 → 58.3 (−4.2)	24.2 → 21.7 (−2.5)	8.3 → 2.5 (−5.8*)	17.5 → 17.5	19.2 → 40.0 (+20.8*)
CE′ (collapsed bonus)	MATH-500	EXACT	1000	16.3 → 15.8 (−0.5)	82.1 → 82.5 (+0.4)	2.7 → 3.8 (+1.1)	1.6 → 0.9 (−0.7)	83.7 → 84.2 (+0.5)
CE′ (collapsed bonus)	MATH-500	NO-SYSTEM	1000	47.7 → 16.4 (−31.3*)	81.5 → 82.5 (+1.0)	31.7 → 3.9 (−27.8*)	28.7 → 1.8 (−26.9*)	38.8 → 80.8 (+42.0*)
CE′ (collapsed bonus)	AIME	EXACT	360	49.4 → 50.0 (+0.6)	27.5 → 27.5	0.3 → 1.4 (+1.1)	11.1 → 10.0 (−1.1)	50.6 → 50.0 (−0.6)
CE′ (collapsed bonus)	AIME	NO-SYSTEM	360	78.9 → 48.3 (−30.6*)	23.3 → 27.5 (+4.2*)	12.8 → 2.2 (−10.6*)	38.1 → 9.2 (−28.9*)	5.0 → 46.9 (+41.9*)
sharpened	MATH-500	EXACT	1000	18.5 → 17.0 (−1.5)	80.8 → 81.6 (+0.8)	3.0 → 2.6 (−0.4)	5.0 → 5.3 (+0.3)	81.5 → 83.0 (+1.5)
sharpened	MATH-500	NO-SYSTEM	1000	47.2 → 20.1 (−27.1*)	79.4 → 81.5 (+2.1)	29.7 → 5.7 (−24.0*)	33.6 → 7.4 (−26.2*)	39.8 → 76.8 (+37.0*)
sharpened	AIME	EXACT	360	57.5 → 57.8 (+0.3)	24.7 → 24.7	1.7 → 1.9 (+0.3)	36.7 → 38.9 (+2.2)	42.5 → 42.2 (−0.3)
sharpened	AIME	NO-SYSTEM	360	82.5 → 58.9 (−23.6*)	21.7 → 25.6 (+3.9*)	11.9 → 1.9 (−10.0*)	59.2 → 31.4 (−27.8*)	8.3 → 41.1 (+32.8*)

A15.2 SHARPENED OBJECTIVE AND MECHANISM VARIANTS

Table 65: Prompt format × objective on the same data, same seed, same budget (MATH-500, 8,192 tokens, $n{=}500$ per cell). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	prompt format	cap	acc	solved-then-capped	loop	well-formed
CE	EXACT	15.6	83.0	2.4	1.0	84.4
CE	HISTORICAL	24.8	79.8	11.6	12.0	56.6
sharpened	EXACT	16.6	82.6	3.4	3.2	83.4
sharpened	HISTORICAL	52.0	76.8	32.2	39.6	31.6

Table 66: Presence versus content of the system turn (Qwen3-8B-Base CE student, seed 42; MATH-500 two decoding seeds; AIME 2025 twelve decoding seeds where available). EMPTY-SYSTEM: a system turn with empty content; JUNK-SYSTEM: a system turn containing an unrelated sentence. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed / unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	cap	acc	solved-then-capped	loop	well-formed	med. tokens
MATH-500, EXACT ($n{=}1000$)	15.7	82.3	2.5	1.3	84.3	1,549
MATH-500, NO-SYSTEM ($n{=}1000$)	30.9	82.2	16.6	15.1	54.1	1,938
MATH-500, EMPTY-SYSTEM ($n{=}1000$)	17.4	80.8	3.7	2.3	74.1	1,386
MATH-500, JUNK-SYSTEM ($n{=}1000$)	21.2	81.6	7.0	6.7	76.6	1,508
MATH-500, ALT-SYSTEM ($n{=}500$)	14.8	84.2	2.8	0.6	85.2	1,468
AIME, EXACT ($n{=}360$)	49.7	26.1	1.1	11.1	50.3	29,638
AIME, NO-SYSTEM ($n{=}360$)	67.2	24.4	8.6	25.0	21.7	30,000
AIME, EMPTY-SYSTEM ($n{=}360$)	53.3	25.3	1.1	12.8	40.3	30,000
AIME, JUNK-SYSTEM ($n{=}120$)	56.7	30.0	5.8	22.5	39.2	30,000
AIME, ALT-SYSTEM ($n{=}360$)	53.9	26.1	1.4	12.2	45.8	30,000
AIME, HISTORICAL ($n{=}360$)	58.3	26.4	6.1	18.3	32.2	30,000
MATH-500, EXACT, 2nd decoding seed, local stack ($n{=}500$)	14.4	83.8	2.2	1.6	85.6	1,508
MATH-500, NO-SYSTEM, 2nd decoding seed, local stack ($n{=}500$)	30.2	83.2	17.2	16.0	54.8	1,933

A15.3 DATA SCALE

Table 67: CE students trained on the same single format with 3 epochs over 30k traces versus 1 epoch over 100k distinct problems (same source distribution; Qwen3-8B-Base; MATH-500 budget 8,192, AIME budget 30,000). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

CE student	bench	format	n	cap	acc	solved-then-capped	loop	well-formed
3 ep × 30k, seed 42	MATH-500	EXACT	1000	15.0	83.4	2.3	1.3	85.0
3 ep × 30k, seed 42	MATH-500	NO-SYSTEM	1000	31.6	82.8	18.4	16.3	54.4
3 ep × 30k, seed 42	AIME	EXACT	120	52.5	29.2	2.5	10.8	47.5
3 ep × 30k, seed 42	AIME	NO-SYSTEM	120	62.5	24.2	8.3	17.5	19.2
3 ep × 30k, seed 123	MATH-500	EXACT	500	16.6	81.0	2.8	1.0	83.4
3 ep × 30k, seed 123	MATH-500	NO-SYSTEM	500	45.6	80.6	29.0	27.8	41.2
3 ep × 30k, seed 123	AIME	EXACT	120	50.8	27.5	0.0	11.7	49.2
3 ep × 30k, seed 123	AIME	NO-SYSTEM	120	68.3	25.8	12.5	28.3	18.3
1 ep × 30k, seed 42	MATH-500	EXACT	1000	19.2	77.4	4.1	3.9	67.6
1 ep × 30k, seed 42	MATH-500	NO-SYSTEM	1000	67.7	78.6	48.0	48.0	1.8
1 ep × 30k, seed 42	AIME	EXACT	120	45.8	20.8	0.8	13.3	44.2
1 ep × 30k, seed 42	AIME	NO-SYSTEM	120	83.3	24.2	15.8	50.0	0.0
1 ep × 100k, seed 42	MATH-500	EXACT	1000	15.8	80.3	1.5	1.5	84.2
1 ep × 100k, seed 42	MATH-500	NO-SYSTEM	1000	49.3	76.5	30.3	31.2	31.8
1 ep × 100k, seed 42	AIME	EXACT	120	50.0	24.2	0.0	9.2	50.0
1 ep × 100k, seed 42	AIME	NO-SYSTEM	120	73.3	27.5	13.3	36.7	10.0
1 ep × 100k, seed 123	MATH-500	EXACT	1000	14.8	82.1	2.2	1.5	85.2
1 ep × 100k, seed 123	MATH-500	NO-SYSTEM	1000	61.2	78.9	43.1	43.0	25.3
1 ep × 100k, seed 123	AIME	EXACT	120	51.7	23.3	0.8	9.2	48.3
1 ep × 100k, seed 123	AIME	NO-SYSTEM	120	86.7	26.7	20.8	48.3	0.8

A15.4 PUBLIC STUDENTS UNDER THEIR OWN AND PERTURBED PROMPT FORMATS

Table 68: Every student in this paper under its own format and one it was not trained on, ordered by lineage. "own → perturbed" is the cap rate; the Δ columns are paired differences. *endpoint*: **yes** if the pre-declared brittleness endpoint is met (cap gap > 5 points and significant); **early** if the student instead fails in the opposite direction, stopping too early or losing accuracy. Cohorts differ by model (budget 8,192 for ours, 16,384 for released students; one or two decoding seeds, n as given), so read each row against itself, not across rows. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

student	initialization	SFT data	n	cap, own → perturbed	Δcap	Δloop	Δprem.	Δacc	endpoint
Qwen3-8B-Base CE (seed 42)	Qwen3-8B-Base	math 30k ×3	500	15.6 → 33.0	+17.4*	+15.6*	+0.6	−0.6	yes
seed 123	Qwen3-8B-Base	math 30k ×3	500	16.6 → 45.6	+29.0*	+26.8*	+0.0	−0.4	yes
1 epoch	Qwen3-8B-Base	math 30k ×1	500	18.8 → 67.0	+48.2*	+43.0*	−0.4	+3.4	yes
science traces	Qwen3-8B-Base	science 30k ×3	500	4.8 → 60.2	+55.4*	+39.8*	−20.8*	+9.0*	yes
code traces	Qwen3-8B-Base	code 30k ×3	500	29.0 → 33.8	+4.8*	+8.0*	−15.0*	+15.4*	no
Qwen2.5-7B CE (seed 42)	Qwen2.5-7B	math 30k ×3	500	21.6 → 58.0	+36.4*	+22.6*	−0.4	−0.2	yes
seed 123	Qwen2.5-7B	math 30k ×3	500	20.4 → 51.2	+30.8*	+20.6*	−0.2	+1.0	yes
DS-R1-Distill-Qwen-32B CE	DS-R1-Distill-32B	math 30k ×3	500	10.0 → 49.6	+39.6*	+16.2*	−0.2	−12.8*	yes
its parent, no SFT	DS-R1-Distill-32B	–	500	5.2 → 7.4	+2.2*	+0.0	−0.6	+4.0*	no
Llama-3.1-8B-Instruct CE	Llama-3.1-8B-Inst	math 30k ×3	500	28.6 → 28.8	+0.2	+1.4	+0.0	−0.6	no
Llama-3.1-8B base CE	Llama-3.1-8B	math 30k ×3	500	33.6 → 36.8	+3.2	+1.2	+0.2	−2.4	no
Mistral-7B-v0.3 CE	Mistral-7B-v0.3	math 30k ×3	500	36.2 → 38.8	+2.6	+0.4	+0.2	−2.0	no
OpenThinker-7B	Qwen2.5-7B-Inst	OpenThoughts 114k	1000	9.1 → 8.1	−1.0	+0.8	+0.5	+0.4	no
DS-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	R1 800k, multi-domain	500	8.2 → 7.0	−1.2	+1.0	+4.0*	−4.2*	no
DS-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	R1 800k, multi-domain	1000	3.1 → 3.7	+0.6	+0.1	+0.1	−3.1*	no
DS-R1-Distill-Qwen-14B	Qwen2.5-14B	R1 800k, multi-domain	500	2.4 → 1.0	−1.4*	−0.2	+0.6	−1.0	no
OpenMath-Nemotron-7B	Qwen2.5-Math-7B	math 3.2M	1000	2.2 → 5.1	+2.9*	+2.1*	+9.6*	−13.9*	early

The parent is flat on the endpoint but not inert. TinyR1-32B-Preview is fine-tuned from DeepSeek-R1-Distill-Qwen-32B, so the two share a family, a template and a size. Under the same added system text the parent does not overrun—its capping moves -0.2 points—but it does not ignore the perturbation either: its median generation falls from 2,210 to 795 tokens, less than half, with a marginal accuracy cost of 4.2 points that sits just above our replication spread. Parent and child therefore move in opposite directions under the same manipulation, one writing far less and

Table 69: The same checkpoint, the same prompts and the same decoding seed, run independently on two clusters (H800 and A800). The paired differences reproduce to within a few tenths of a point, which bounds how much of any small effect in Table 71 could be run-to-run noise.

	Δcap			Δacc		
model	run A	run B	|diff|	run A	run B	|diff|
Qwen3.8-27B	−1.4	−2.2	0.8	+17.8	+14.8	3.0
QwQ-32B-Preview	+0.2	+0.2	0.0	−0.6	−2.2	1.6
Light-R1-32B	−0.4	+0.2	0.6	+1.2	+0.6	0.6
TinyR1-32B-Preview	+11.0	+12.0	1.0	+0.0	−0.8	0.8
s1-32B	−0.2	+0.6	0.8	−0.8	−4.2	3.4
s1.1-32B	+2.0	+1.2	0.8	+0.2	−0.6	0.8
Nemotron-3-Nano-30B-A3B	−2.6	−2.2	0.4	+1.0	+0.8	0.2
gpt-oss-20b	+1.8	+1.6	0.2	−11.2	−10.0	1.2

Table 70: Of 31 released checkpoints, the configuration our students fail under—built from a raw base *and* served with a system turn that can be removed—occurs 4 times. Collapsing the lower two rows, base-initialized models carry a system turn in 4/20 cases against 7/11 for already-post-trained ones (Fisher exact $p = 0.023$).

fine-tuned from	serves a system turn	does not	total
base LM	4	16	20
instruct checkpoint	7	0	7
another reasoning model	0	4	4
total	11	20	31

the other writing more and failing to stop. Because the two native cells already sit at opposite poles, a crossover between them is unusually informative, and we pre-registered one: the parent's prompt with a thought opener prefilled, and the child's with it removed. The formats were verified byte-identical along each diagonal before any generation existed, so on each diagonal only the weights differ. All four readings, including the two that would kill the hypothesis, were fixed in advance.

Why adding a turn can break a model that never had one. We first offered one account covering both directions—the start of the sequence does not look like training did—and pre-registered it as the prediction for the mirror arm of Table 25. It failed: our own student trained with no system turn does not break when a generic one is added ($+1.4$ $[-0.5, +3.3]$), nor under an unrelated or an empty turn, and its closing margin loses 0.38 of its value where the main student's loses 0.90 under deletion (Table 27). The same table holds nineteen further models, sixteen of them base-initialized, for which inserting a turn where the template has none never raises capping. So the addition direction is not the mirror image of deletion, and TinyR1's sensitivity to an inserted turn—raw undelimited text between the beginning-of-sequence and user markers in its template—must rest on something its training prefix contains and ours does not; the pre-registered crossover on it and its own parent checkpoint is reported in Table 29. Its break is not a margin-share effect: pre-registered, TinyR1's end-of-turn share under the added turn is 0.11 (95% CI $[0.03, 0.16]$), inside the holding range, against -0.04 for its parent; what differs is the size of the margin ($+4.9$ against $+15.9$), so its route stays untested.

How rare the configuration is. Across 31 released reasoning checkpoints the two are close to mutually exclusive: of the 20 built from a raw base only 4 are served with a system turn, while all 7 built from an instruct checkpoint are and none of the 4 built from another reasoning model is, that family's template having no system slot (Fisher exact p=0.023; Table 70). Those four are the whole clean test set. Two place a reasoning directive in the turn, so deleting it also switches reasoning behaviour. One, Nemotron-3-Nano-30B-A3B with an empty block, moves in the protective direction (-2.6 points, significant) and is what falsifies raw-base initialization as a sufficient cause; it passes our marker criterion for a long-CoT model but not our median-length one, so we report it under both rules rather than choosing. The fourth, OpenMath-Nemotron-7B, is the one released student this perturbation damages (its cells under six formats: Table 72).

Table 71: Released long-CoT checkpoints under their own documented prompt format and under a perturbation of the system turn. MATH-500, budget 16,384. Intervals are item-clustered bootstrap 95% CIs on the paired difference.

released model	init	sys	sys kind	dom	n	cap own	Δcap [95% CI]	Δacc	Δprem	LC
OpenMath-Nemotron-7B	base	yes	empty	1	1000	2.2	$+2.9$ [+1.5, +4.4]	-13.9	$+9.6$	Y/Y
s1.1-32B	inst	yes	generic	2	500	3.2	$+2.0$ [+0.4, +3.8]	$+0.2$	$+0.0$	Y/Y
gpt-oss-20b	base	yes	reasoning	4	500	1.8	$+1.8$ [+0.0, +3.6]	-11.2	$+10.4$	N/Y
QwQ-32B-Preview	inst	yes	task	2	500	1.2	$+0.2$ [-1.0, +1.6]	-0.6	$+0.0$	Y/N
s1-32B	inst	yes	generic	2	500	0.4	-0.2 [-1.0, +0.4]	-0.8	-0.6	Y/Y
Light-R1-32B	inst	yes	generic	1	500	3.0	-0.4 [-2.0, +1.0]	$+1.2$	-0.4	Y/Y
Llama-3.1-Nemotron-Nano-8B	inst	yes	reasoning	5	500	2.0	-0.4 [-1.8, +1.0]	$+0.6$	$+0.6$	Y/Y
OpenReasoning-Nemotron-32B	inst	yes	empty	3	500	7.2	-0.4 [-2.0, +1.4]	$+0.6$	$+0.0$	Y/Y
OpenThinker-7B	inst	yes	task	3	1000	9.1	-1.0 [-2.9, +0.9]	$+0.4$	$+0.5$	Y/Y
Qwen3.8-27B	base	yes	reasoning	4	500	2.0	-1.4 [-2.8, -0.2]	$+17.8$	-16.8	N/Y
Nemotron-3-Nano-30B-A3B	base	yes	empty	4	500	5.2	-2.6 [-4.4, -1.0]	$+1.0$	$+0.2$	N/Y
TinyR1-32B-Preview	reas	no	–	3	500	7.2	$+11.0$ [+7.4, +14.6]	$+0.0$	$+0.0$	Y/Y
Qwen3-32B	base	no	–	4	500	2.8	$+0.6$ [-1.0, +2.0]	-0.2	$+0.2$	Y/Y
DS-R1-Distill-Qwen-7B	base	no	–	4	1000	3.1	$+0.6$ [-0.5, +1.8]	-3.1	$+0.1$	Y/Y
Qwen3-8B	base	no	–	4	500	4.4	$+0.0$ [-2.0, +2.0]	-0.6	$+0.4$	Y/Y
Qwen3-14B	base	no	–	4	500	3.0	$+0.0$ [-1.4, +1.4]	$+0.6$	$+0.2$	Y/Y
Light-R1-7B-DS	reas	no	–	1	500	2.8	$+0.0$ [-1.4, +1.4]	-3.4	$+2.8$	Y/Y
OpenMath2-Llama3.1-8B	base	no	–	1	500	0.4	$+0.0$ [-0.8, +0.8]	$+1.6$	-0.4	N/N
DS-R1-Distill-Qwen-32B	base	no	–	4	500	2.0	-0.2 [-1.8, +1.4]	-4.2	$+0.6$	Y/Y
Light-R1-32B-DS	reas	no	–	1	500	1.8	-0.4 [-1.8, +1.0]	-1.2	-0.2	Y/Y
Seed-OSS-36B-Instruct	base	no	–	4	500	5.4	-0.6 [-2.2, +0.8]	$+0.0$	$+0.2$	Y/Y
Qwen3-30B-A3B-Thinking	base	no	–	4	500	4.4	-0.8 [-2.2, +0.8]	$+0.8$	$+0.0$	Y/Y
QwQ-32B	base	no	–	3	500	2.8	-0.8 [-2.2, +0.6]	-0.4	$+1.0$	Y/Y
Light-R1-14B-DS	reas	no	–	1	500	3.0	-1.2 [-2.6, +0.2]	$+0.6$	$+0.2$	Y/Y
DS-R1-Distill-Qwen-1.5B	base	no	–	4	500	8.2	-1.2 [-3.6, +1.2]	-4.2	$+4.0$	Y/Y
DS-R1-Distill-Qwen-14B	base	no	–	4	500	2.4	-1.4 [-2.8, -0.2]	-1.0	$+0.6$	Y/Y
Qwen3.5-9B	base	no	–	4	500	31.4	-1.6 [-6.0, +3.0]	$+0.2$	$+1.4$	Y/Y
DS-R1-0528-Qwen3-8B	base	no	–	4	500	11.2	-1.6 [-3.8, +0.6]	-1.6	$+3.2$	Y/Y
Qwen3-4B-Thinking-2507	base	no	–	4	500	8.4	-2.0 [-3.8, -0.2]	$+1.2$	$+0.4$	Y/Y
Qwen3.5-27B	base	no	–	4	500	20.8	-4.6 [-8.8, -0.6]	$+0.6$	$+4.0$	Y/Y
Qwen3.6-27B	base	no	–	4	500	21.4	-6.0 [-8.8, -3.4]	$+13.2$	-7.2	Y/Y

We looked for a predictor, and what we found was three falsifications. Each explanation was pre-registered before the models that would test it had produced a single row, with the zero-row state timestamped so a result could not influence whether it counted. All three fell. *Few training steps* predicts that a smaller dose breaks harder, since our own one-epoch student has the largest gap of any; s1-32B, at $1{,}000$ fine-tuning examples the smallest dose in the panel, does not move. *Single-domain post-training*: Light-R1-32B, fine-tuned from an instruct checkpoint on single-domain mathematics with a generic system turn, does not move at full sample ($+0.2$ points $[-1.2, +1.6]$). Both are fine-tuned from instruct checkpoints that have already seen many formats, so neither is a clean test for the base-initialized case our own students occupy. *Raw-base initialization* looked perfectly separating on the four models available at the time; Nemotron-3-Nano-30B-A3B is base-initialized with an empty system block—a purely structural deletion with nothing in it to confound—and moves -2.6 points of capping, in the protective direction.

No single property survives. The one released student that is damaged satisfies two of them at once: it is a raw-base student post-trained on a single domain in a single format, which is also what our own brittle students are, and what the DeepSeek-R1-Distill checkpoints are before we fine-tune them further. A conjunction is therefore the only form not yet falsified, which is not the same as a form that has been confirmed—it rests on one positive case, two of the three falsifications are at half their planned sample, and one depends on the more inclusive of our two long-CoT criteria. The two negatives carry whatever weight there is here; the single positive carries none. We give the numbers under both criteria and leave the question open.

The released panel fails the other way round, when it fails at all. Across the fourteen released long-CoT students with a paired arm, the change in accuracy is as often positive as negative—seven each way, sign test $p = 1.00$—so there is no systematic accuracy drift under the perturbation and we do not claim one. What is true is narrower. Under deletion exactly one model shows a significant increase in capping (OpenMath-Nemotron-7B, $+2.9$ points, still below our endpoint; s1.1-32B's $+2.0$ is a second, in the full panel of Table 71), and under addition only TinyR1 does. The accuracy channel needs a noise floor before anything is read from it. On the eight models that ran twice (Table 69), the paired change in capping reproduces to within 1.0 point but the paired change in

accuracy only to within 3.4, so an accuracy effect smaller than that is not distinguishable from run-to-run variation. Three effects clear it. Two are damage: OpenMath-Nemotron-7B loses 13.9 points and gpt-oss-20b loses 11.2 and 10.0 on two independent runs. One is the opposite: Qwen3.8-27B *gains* 17.8 points when its system turn is deleted—but that turn contains a reasoning directive, so what improves it is switching reasoning mode, not the format change. The remaining accuracy movements sit inside the replication spread and we draw nothing from them.

The shape of that minority is the mirror image of our own students. Ours are damaged by failing to stop: they write the answer and run to the budget. The released students that are damaged fail by stopping too soon, losing accuracy and ending the turn with no answer written. The same perturbation, applied to models trained under different recipes, moves the stop decision in opposite directions.

Two hazards in scoring a model you did not train. First, the thought-closing marker is not universal, and assuming it is turns a healthy model into a broken one. Of gpt-oss-20b's 500 MATH-500 generations, 491 stop naturally and 493 contain its own closer, `<|channel|>final`; neither `</think>` nor `<|end_of_thought|>` appears even once. Scored against the two markers our own students use, it would read as 0% well-formed stops—catastrophically broken—when 98% of its generations close correctly. We have now met three such models, one of them predicted in advance from its template before it was run. We therefore record each released model's own closer before scoring it, and the capping rate, which needs no marker, is what the panel rests on.

Second, deletion is not one experiment either, because system turns differ in kind. Across the released panel they fall into four classes: an *empty* block carrying structure and no text at all; a *generic* persona line, which is what our own students carry and therefore the only clean analogue of our perturbation; a *task-framing* instruction; and a *reasoning-mode directive*. A system turn may thus carry more than format. The harmony preamble of gpt-oss-20b contains `Reasoning: medium` and Llama-3.1-Nemotron-Nano-8B's contains `detailed thinking on`, so deleting the system turn from those two removes a reasoning-effort directive along with the serialization. For them the perturbation is not format-only, and their flatness should be read with that caveat rather than as evidence that format does not matter. These two facts interact badly for anyone hoping to settle the question with more released models: the checkpoints that would separate initialization from training domain within the clean empty-block class either turn out not to be long-CoT under their documented call, or carry a reasoning directive in the very system turn the perturbation deletes. An empty system block may itself mean reduced reasoning effort for families that put the toggle there. We therefore expect this predictor to be difficult to settle from released checkpoints at all, and say nothing about it. We also excluded gpt-oss-20b from the panel on separate grounds: under its documented default call its MATH-500 generations run a median 713 tokens against 2,100 to 9,000 for the rest, so it is not a long-CoT model in that configuration.

Harm the termination metric does not see. The perturbation can damage a released student without moving the outcome this paper leads with. Adding a generic system turn to DeepSeek-R1-0528-Qwen3-8B leaves capping unchanged (-1.6 $[-3.8, +0.6]$) and accuracy within noise (-1.6 $[-4.4, +1.2]$) while adding 3.2 $[+1.8, +4.8]$ points of premature stops; a termination-only reading of that model reports nothing at all. Two cautions follow. A difference in capping should always be read beside the model's own-format rate, which ranges from 2 to 31% across the released students before any perturbation—Qwen3.5-9B already caps 31.4% under its own format, so it has far less headroom than a student that caps 2%. And an evaluation that watches only termination is exposed to the same blindness we criticise in evaluations that watch only accuracy.

Table 72: Released long-CoT students evaluated under their training format and perturbations of it (MATH-500 budget 16,384, two decoding seeds; AIME 2025 budget 30,000, four seeds). OpenThinker-7B: training format = the OpenThoughts system prompt + the user prefix "Return your final response within `\boxed{}`."; "template default" is what its chat template produces when no system message is supplied. DeepSeek-R1-Distill-Qwen-7B: training format has no system turn; the "recommended sentence" is the model card's "Please reason step by step, and put your final answer within `\boxed{}`." OpenMath-Nemotron-7B: training format = empty system block + the NeMo-Skills instruction line; "no system block" removes the block that its chat template inserts (without block and instruction line, 686 of its 1,000 MATH-500 generations are a single end-of-text token). Premature = natural stop without a boxed answer. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

public student	format	bench	n	cap	acc	solved-then-capped	loop	premature	med. tokens
OpenThinker-7B	training format (OpenThoughts system + prefix)	MATH-500	1000	9.1	82.5	0.1	0.7	3.2	2,547
	training format (OpenThoughts system + prefix)	AIME	120	15.0	21.7	0.0	12.5	1.7	16,564
	template default (Qwen system) + prefix	MATH-500	1000	6.6	83.0	0.0	0.6	3.7	2,583
	template default (Qwen system) + prefix	AIME	120	13.3	29.2	0.0	9.2	0.0	15,655
	generic system + prefix	MATH-500	1000	7.9	82.4	0.3	0.5	3.9	2,533
	generic system + prefix	AIME	120	14.2	26.7	0.0	10.0	0.8	17,472
	no system + prefix	MATH-500	1000	8.1	82.9	0.3	1.5	3.7	2,503
	no system + prefix	AIME	120	15.8	27.5	0.0	10.8	0.0	16,297
	training system, bare problem	MATH-500	1000	8.2	79.1	0.0	0.7	5.4	2,491
	training system, bare problem	AIME	120	15.8	25.8	0.0	12.5	0.8	17,070
DeepSeek-R1-Distill-Qwen-7B	no system, bare problem (native)	MATH-500	1000	3.1	91.7	0.1	0.0	1.0	2,374
	no system, bare problem (native)	AIME	120	15.8	40.8	0.0	1.7	0.8	11,219
	no system + recommended sentence	MATH-500	1000	2.7	92.0	0.0	0.1	1.0	2,324
	no system + recommended sentence	AIME	120	8.3	39.2	0.0	1.7	0.8	13,815
	generic system, bare problem	MATH-500	1000	3.7	88.6	0.1	0.1	1.1	1,693
	generic system, bare problem	AIME	120	10.0	38.3	0.0	2.5	0.0	12,721
	generic system + recommended sentence	MATH-500	1000	3.0	91.7	0.0	0.0	1.1	2,162
	generic system + recommended sentence	AIME	120	9.2	42.5	0.0	0.8	0.8	13,123
OpenMath-Nemotron-7B	training format (empty system block + instruction)	MATH-500	1000	2.2	93.8	0.2	0.0	1.2	2,671
	training format (empty system block + instruction)	AIME	120	1.7	67.5	0.0	0.0	0.0	14,806
	no system block + instruction	MATH-500	1000	5.1	79.9	0.6	2.1	10.8	2,568
	no system block + instruction	AIME	120	2.5	55.0	0.0	0.8	5.0	14,177
	empty system block, bare problem	MATH-500	1000	2.2	93.4	0.5	0.0	1.0	2,582
	empty system block, bare problem	AIME	120	0.8	60.8	0.0	0.8	0.0	13,806
	no system, bare problem	MATH-500	1000	2.2	18.3	0.1	0.6	71.7	1
	no system, bare problem	AIME	120	0.8	18.3	0.0	0.0	56.7	1
	generic system + instruction	MATH-500	1000	2.8	94.1	0.3	0.0	1.0	2,489
	generic system + instruction	AIME	120	1.7	61.7	0.0	0.0	0.0	14,672
	Qwen default system + instruction	MATH-500	1000	3.0	93.4	0.1	0.0	1.0	2,472
	Qwen default system + instruction	AIME	120	1.7	62.5	0.0	0.0	0.8	14,107
OpenReasoning-Nemotron-32B	no system block + instruction	MATH-500	500	6.8	91.8	0.4	0.0	0.4	3,960
	generic system + instruction	MATH-500	500	6.6	92.2	0.6	0.0	0.6	3,981
	no system, bare problem	MATH-500	500	7.2	91.2	0.0	0.0	0.8	3,893
DeepSeek-R1-Distill-Qwen-14B	no system, bare problem (native)	MATH-500	500	2.4	91.8	0.4	0.2	1.2	2,219
	generic system, bare problem	MATH-500	500	1.0	90.8	0.2	0.0	1.8	1,836
	no system + instruction	MATH-500	500	1.4	92.8	0.0	0.0	1.2	1,569
DeepSeek-R1-Distill-Qwen-1.5B	no system, bare problem (native)	MATH-500	500	8.2	80.8	0.0	0.6	1.6	2,474
	generic system, bare problem	MATH-500	500	7.0	76.6	0.2	1.6	5.6	1,537
	no system + instruction	MATH-500	500	7.4	78.2	0.2	0.6	1.2	1,332
OpenMath2-Llama3.1-8B (not long-CoT)	training format (no system + instruction)	MATH-500	500	0.4	64.6	0.0	0.4	1.8	265
	generic system + instruction	MATH-500	500	0.4	66.2	0.0	0.4	1.4	265
	no system, bare problem	MATH-500	500	0.2	64.6	0.0	0.4	1.2	262

A16 REPAIRS

A16.1 AUGMENTATION-FRACTION SWEEP

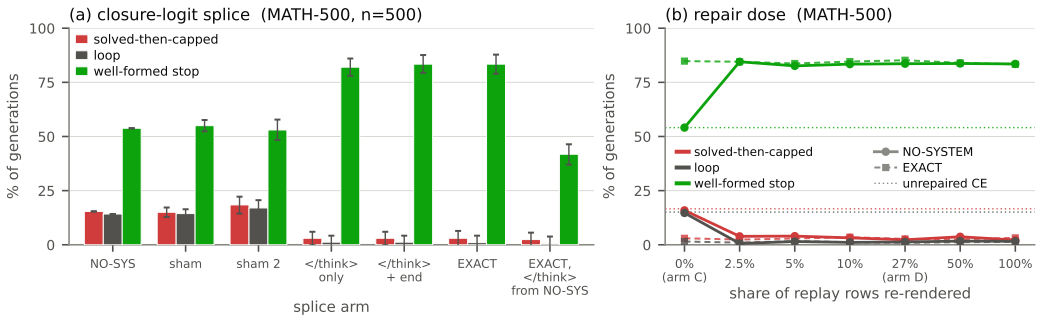

Figure 11: **Left:** the closure-logit splice of Table 56: outcome rates under NO-SYSTEM sampling when the named tokens' probabilities are taken from the EXACT conditional, and (last group) under EXACT sampling with `</think>` taken from NO-SYSTEM (error bars: 95% CI of the paired difference to plain NO-SYSTEM sampling). **Right:** the augmentation-fraction sweep of Table 73: NO-SYSTEM outcomes (solid) and EXACT outcomes (dashed) of one-epoch repairs that reformat the given share of replay rows; dotted lines are the unrepaired CE student under NO-SYSTEM.

Table 73: Fraction of the 3,420 replay rows reformatted under NO-SYSTEM / HISTORICAL (equal split; nested subsets), one epoch from the CE student, one training run per fraction. 0% is arm C and 27% is arm D (seed-42 rows re-evaluated on the cluster stack); the other rows are new runs. Seed 42: MATH-500 EXACT and NO-SYSTEM pool two decoding seeds (n=1000), held-out formats one (n=500), AIME four (n=120). Seed 123 rows: MATH-500 one decoding seed (n=500), AIME four. Values in %; Figure 11 (right) plots the sweep. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

reformatted	MATH-500 NO-SYSTEM: cap	solved-then-capped	loop	wf	acc	EXACT: cap	loop	acc	ALT-SYS loop	NO-SYS+ALT-INSTR loop	AIME NO-SYSTEM: cap	loop	wf	AIME EXACT cap
0% (arm C)	32.8	15.9	14.7	54.1	79.8	15.2	1.5	83.8	0.8	14.4	69.2	22.5	14.2	50.0
2.5%	15.4	3.9	0.6	84.5	83.8	15.5	1.0	82.3	1.4	1.2	55.0	10.0	45.0	50.0
2.5%, 2nd draw	17.2	2.7	1.5	82.4	81.6	15.5	0.8	81.1	2.4	1.0	47.5	9.2	52.5	55.0
2.5%, 3rd draw	16.4	3.1	0.7	83.5	82.1	16.7	1.6	83.1	1.0	1.2	53.3	13.3	46.7	55.8
5%	17.1	4.0	1.5	82.6	82.6	16.3	1.7	82.4	1.2	0.8	54.2	6.7	44.2	49.2
10%	16.6	3.2	1.1	83.4	82.3	15.4	1.2	83.7	1.6	1.8	48.3	10.0	51.7	50.8
27% (arm D)	16.4	2.3	1.3	83.6	81.4	14.8	0.9	83.2	1.0	1.6	50.0	14.2	50.0	48.3
50%	16.3	3.7	1.7	83.7	82.3	16.0	1.2	81.7	2.2	1.4	52.5	16.7	47.5	53.3
100%	16.5	2.2	1.6	83.5	80.9	16.7	1.3	82.9	1.4	0.6	51.7	10.0	48.3	45.8
0% (arm C) [seed 123]	41.2	26.4	25.2	43.0	80.2	17.2	1.2	81.4	1.6	36.8	68.3	31.7	16.7	48.3
2.5% [seed 123]	15.4	3.0	1.6	84.6	83.2	16.5	1.2	82.8	1.6	2.0	55.0	15.8	43.3	52.5
10% [seed 123]	16.8	4.1	1.4	83.0	82.9	15.6	0.6	81.7	1.6	1.4	50.8	14.2	48.3	55.0
27% (arm D) [seed 123]	16.8	3.2	2.0	83.2	81.4	17.4	0.6	82.0	0.8	1.4	52.5	10.0	47.5	48.3

A16.2 THE REPAIR OUTSIDE MATHEMATICS

Arms C and D rebuilt from the science student's own training set exactly as the math arms were: 3,420 rows, of which 912 (27%) are reformatted under NO-SYSTEM or HISTORICAL in arm D and none in arm C; one epoch from its checkpoint, LR 2×10^{-6} constant after 5% warm-up, full loss. Each row of Table 74 gives the paired difference between that model's own EXACT and NO-SYSTEM cells, so a smaller number is a smaller brittleness gap.

Table 74: The science student and its two repairs, MATH-500 (one decoding seed) and GPQA-Diamond (four seeds), each under EXACT and NO-SYSTEM. The unrepaired rows are a separate decode of the same checkpoint on the arms' own runtime, so they differ from Tables 47 and 68 by run-to-run noise (≤ 1.4 points on MATH-500, ≤ 2.5 on GPQA; the MATH-500 gap is $+55.4$ in both). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

model	bench	n	cap EXACT	cap NO-SYSTEM	loop EXACT	loop NO-SYSTEM	wf EXACT	wf NO-SYSTEM	acc EXACT / NO-SYSTEM
science student (unrepaired)	MATH-500	500	5.2	60.6 (+55.4*)	1.8	43.0 (+41.2*)	94.8	29.4 (−65.4*)	47.6 / 58.0
	GPQA-Diamond	792	15.0	63.3 (+48.2*)	15.2	53.9 (+38.8*)	69.3	14.5 (−54.8*)	0.3 / 0.4
+ arm C, replay only	MATH-500	500	5.6	57.0 (+51.4*)	2.8	36.8 (+34.0*)	94.4	34.2 (−60.2*)	42.8 / 61.0
	GPQA-Diamond	792	16.4	55.1 (+38.6*)	16.4	44.9 (+28.5*)	70.5	19.6 (−50.9*)	0.1 / 0.5
+ arm D, format-augmented replay	MATH-500	500	5.2	12.0 (+6.8*)	2.0	6.6 (+4.6*)	94.8	88.0 (−6.8*)	42.6 / 48.0
	GPQA-Diamond	792	16.9	28.8 (+11.9*)	16.8	27.8 (+11.0*)	67.6	55.3 (−12.2*)	0.0 / 0.4

A16.3 REPAIR REPLICATIONS: SECOND TRAINING SEED AND 32B STUDENT

Table 75: Repair arms B, C, D trained from the seed-123 CE student, evaluated against that student (paired; MATH-500 one decoding seed, n=500; AIME four seeds, n=120). Same conventions as Table 77. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	model	cap	acc	solved-then-capped	loop	well-formed	premature
EXACT	reference	16.6	81.0	2.8	1.0	83.4	0.6
	B teacher-prefix terminal	17.4 (+0.8)	81.0	3.4 (+0.6)	3.4 (+2.4*)	82.6 (−0.8)	0.6
	C replay only	17.2 (+0.6)	81.4 (+0.4)	3.0 (+0.2)	1.2 (+0.2)	82.8 (−0.6)	0.8 (+0.2)
	D format-augmented replay	17.4 (+0.8)	82.0 (+1.0)	3.0 (+0.2)	0.6 (−0.4)	82.6 (−0.8)	0.6
NO-SYSTEM	reference	45.6	80.6	29.0	27.8	41.2	0.6
	B teacher-prefix terminal	17.0 (−28.6*)	83.2 (+2.6)	5.4 (−23.6*)	3.4 (−24.4*)	79.2 (+38.0*)	0.4 (−0.2)
	C replay only	41.2 (−4.4)	80.2 (−0.4)	26.4 (−2.6)	25.2 (−2.6)	43.0 (+1.8)	0.8 (+0.2)
	D format-augmented replay	16.8 (−28.8*)	81.4 (+0.8)	3.2 (−25.8*)	2.0 (−25.8*)	83.2 (+42.0*)	0.4 (−0.2)
ALT-SYSTEM (held out)	reference	16.8	80.4	2.2	1.0	83.0	0.6
	B teacher-prefix terminal	15.8 (−1.0)	82.6 (+2.2)	3.6 (+1.4)	2.6 (+1.6)	84.0 (+1.0)	0.6
	C replay only	16.8	81.2 (+0.8)	3.0 (+0.8)	1.6 (+0.6)	82.8 (−0.2)	0.6
	D format-augmented replay	15.0 (−1.8)	82.8 (+2.4)	2.6 (+0.4)	0.8 (−0.2)	85.0 (+2.0)	0.6
NO-SYSTEM+ALT-INSTR (held out)	reference	45.8	72.2	22.8	31.6	34.4	0.8
	B teacher-prefix terminal	13.6 (−32.2*)	82.0 (+9.8*)	2.0 (−20.8*)	0.8 (−30.8*)	84.2 (+49.8*)	0.6 (−0.2)
	C replay only	53.4 (+7.6*)	73.6 (+1.4)	31.0 (+8.2*)	36.8 (+5.2*)	29.8 (−4.6)	0.8
	D format-augmented replay	15.2 (−30.6*)	83.6 (+11.4*)	3.0 (−19.8*)	1.4 (−30.2*)	84.8 (+50.4*)	0.4 (−0.4)
AIME EXACT	reference	50.8	27.5	0.0	11.7	49.2	0.0
	B teacher-prefix terminal	50.8	24.2 (−3.3)	1.7 (+1.7)	17.5 (+5.8)	49.2	0.0
	C replay only	48.3 (−2.5)	27.5	0.0	17.5 (+5.8)	51.7 (+2.5)	0.0
	D format-augmented replay	48.3 (−2.5)	27.5	1.7 (+1.7)	10.8 (−0.8)	51.7 (+2.5)	0.0
AIME NO-SYSTEM	reference	68.3	25.8	12.5	28.3	18.3	0.0
	B teacher-prefix terminal	46.7 (−21.7*)	25.8	0.8 (−11.7*)	10.0 (−18.3*)	51.7 (+33.3*)	1.7 (+1.7)
	C replay only	68.3	26.7 (+0.8)	10.0 (−2.5)	31.7 (+3.3)	16.7 (−1.7)	0.0
	D format-augmented replay	52.5 (−15.8*)	25.8	0.0 (−12.5*)	10.0 (−18.3*)	47.5 (+29.2*)	0.0

Table 76: Arm D trained from the DeepSeek-R1-Distill-Qwen-32B CE student (8 GPUs, ZeRO-3, packing length 24,576 which drops the 0.8% of rows longer than that; otherwise the 8B recipe), evaluated against the same student on the same runtime stack (MATH-500 one decoding seed, n=500; AIME four seeds, n=120). Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	model	cap	acc	solved-then-capped	loop	well-formed	premature
EXACT	reference	10.6	89.8	3.6	2.2	89.4	0.6
	D format-augmented replay	12.4 (+1.8)	88.4 (−1.4)	3.2 (−0.4)	0.8 (−1.4)	87.6 (−1.8)	0.4 (−0.2)
NO-SYSTEM	reference	49.6	79.4	30.2	17.4	50.4	0.2
	D format-augmented replay	11.8 (−37.8*)	89.6 (+10.2*)	4.4 (−25.8*)	1.4 (−16.0*)	88.2 (+37.8*)	1.0 (+0.8*)
HISTORICAL	reference	14.0	90.6	6.6	4.6	86.0	0.2
	D format-augmented replay	11.2 (−2.8)	89.0 (−1.6)	3.4 (−3.2*)	1.4 (−3.2*)	88.8 (+2.8)	0.2
ALT-SYSTEM	reference	19.8	88.4	10.2	6.6	80.2	0.4
	D format-augmented replay	14.8 (−5.0*)	88.2 (−0.2)	6.0 (−4.2*)	3.2 (−3.4*)	85.2 (+5.0*)	0.4
NO-SYSTEM+ALT-INSTR	reference	40.6	82.2	24.6	11.0	59.4	0.8
	D format-augmented replay	13.2 (−27.4*)	90.4 (+8.2*)	6.0 (−18.6*)	1.8 (−9.2*)	86.8 (+27.4*)	0.6 (−0.2)
AIME EXACT	reference	33.3	61.7	3.3	2.5	66.7	0.0
	D format-augmented replay	36.7 (+3.3)	61.7	5.8 (+2.5)	5.8 (+3.3)	63.3 (−3.3)	0.0
AIME NO-SYSTEM	reference	64.2	41.7	10.0	30.0	35.8	0.0
	D format-augmented replay	30.8 (−33.3*)	61.7 (+20.0*)	1.7 (−8.3*)	4.2 (−25.8*)	69.2 (+33.3*)	0.0

A16.4 REPAIR ARMS: MATH-500 AND AIME 2025 TABLES

Table 77: Repair arms vs. the CE student they start from, MATH-500, budget 8,192. Values are percentages; parentheses give the paired difference to the CE student (* significant). EXACT and NO-SYSTEM pool two decoding seeds (n=1000); the two held-out prompt formats, used by none of the arms, have one seed (n=500); AIME 2025 in Table 78.

prompt format	model	cap	acc	solved-then-capped	loop	well-formed	premature
EXACT	CE student	15.7	82.3	2.5	1.3	84.3	0.5
	A on-policy terminal	14.5 (−1.2)	82.1 (−0.2)	1.8 (−0.7)	0.6 (−0.7)	85.5 (+1.2)	0.9 (+0.4)
	A, full-sequence loss	37.6 (+21.9*)	79.7 (−2.6*)	21.0 (+18.5*)	26.5 (+25.2*)	62.4 (−21.9*)	0.2 (−0.3)
	B teacher-prefix terminal	17.8 (+2.1)	81.9 (−0.4)	4.4 (+1.9*)	2.1 (+0.8)	82.2 (−2.1)	0.7 (+0.2)
	B, full-sequence loss	15.0 (−0.7)	82.1 (−0.2)	2.0 (−0.5)	1.6 (+0.3)	85.0 (+0.7)	0.4 (−0.1)
	C replay only	15.2 (−0.5)	83.8 (+1.5)	3.0 (+0.5)	1.5 (+0.2)	84.8 (+0.5)	0.5
	D format-augmented replay	14.8 (−0.9)	83.2 (+0.9)	2.7 (+0.2)	0.9 (−0.4)	85.2 (+0.9)	0.7 (+0.2)
NO-SYSTEM	CE student	30.9	82.2	16.6	15.1	54.1	0.5
	A on-policy terminal	15.9 (−15.0*)	81.8 (−0.4)	3.6 (−13.0*)	1.3 (−13.8*)	72.9 (+18.8*)	1.9 (+1.4*)
	A, full-sequence loss	34.3 (+3.4)	82.3 (+0.1)	19.5 (+2.9)	22.7 (+7.6*)	65.0 (+10.9*)	0.5
	B teacher-prefix terminal	16.9 (−14.0*)	83.1 (+0.9)	4.5 (−12.1*)	3.2 (−11.9*)	80.1 (+26.0*)	0.8 (+0.3)
	B, full-sequence loss	16.0 (−14.9*)	83.0 (+0.8)	3.1 (−13.5*)	1.6 (−13.5*)	83.0 (+28.9*)	0.6 (+0.1)
	C replay only	32.8 (+1.9)	79.8 (−2.4*)	15.9 (−0.7)	14.7 (−0.4)	54.1	0.4 (−0.1)
	D format-augmented replay	16.4 (−14.5*)	81.4 (−0.8)	2.3 (−14.3*)	1.3 (−13.8*)	83.6 (+29.5*)	0.8 (+0.3)
ALT-SYSTEM (held out)	CE student	14.8	84.2	2.8	0.6	85.2	0.8
	A on-policy terminal	15.6 (+0.8)	83.8 (−0.4)	4.0 (+1.2)	0.8 (+0.2)	76.4 (−8.8*)	0.2 (−0.6)
	A, full-sequence loss	37.0 (+22.2*)	79.4 (−4.8*)	19.0 (+16.2*)	25.0 (+24.4*)	63.0 (−22.2*)	0.0 (−0.8*)
	B teacher-prefix terminal	18.0 (+3.2*)	81.6 (−2.6)	4.2 (+1.4)	2.8 (+2.2*)	81.4 (−3.8*)	0.6 (−0.2)
	B, full-sequence loss	16.2 (+1.4)	80.8 (−3.4*)	3.0 (+0.2)	1.4 (+0.8)	83.6 (−1.6)	0.4 (−0.4)
	C replay only	15.0 (+0.2)	84.0 (−0.2)	2.8	0.8 (+0.2)	84.8 (−0.4)	0.8
	D format-augmented replay	15.2 (+0.4)	83.4 (−0.8)	3.2 (+0.4)	1.0 (+0.4)	84.8 (−0.4)	0.4 (−0.4)
NO-SYSTEM+ALT-INSTR (held out)	CE student	29.2	79.2	13.8	15.4	55.8	0.6
	A on-policy terminal	12.0 (−17.2*)	83.0 (+3.8*)	2.2 (−11.6*)	1.0 (−14.4*)	86.8 (+31.0*)	1.0 (+0.4)
	A, full-sequence loss	35.0 (+5.8*)	80.0 (+0.8)	17.4 (+3.6)	23.0 (+7.6*)	63.6 (+7.8*)	0.6
	B teacher-prefix terminal	16.6 (−12.6*)	80.2 (+1.0)	3.2 (−10.6*)	2.8 (−12.6*)	80.8 (+25.0*)	0.8 (+0.2)
	B, full-sequence loss	16.2 (−13.0*)	82.0 (+2.8)	4.2 (−9.6*)	1.4 (−14.0*)	82.6 (+26.8*)	0.8 (+0.2)
	C replay only	30.4 (+1.2)	78.6 (−0.6)	14.0 (+0.2)	14.4 (−1.0)	57.4 (+1.6)	0.4 (−0.2)
	D format-augmented replay	15.6 (−13.6*)	83.2 (+4.0*)	3.0 (−10.8*)	1.6 (−13.8*)	84.4 (+28.6*)	0.8 (+0.2)

Table 78: Repair arms vs. the CE student, AIME 2025, budget 30,000, twelve decoding seeds (n=360, paired, same runtime stack). Same conventions as Table 77. Values are percentages of generations. *cap*: reached the token budget; solved-then-capped: capped although the last boxed answer was correct; *well-formed* / *unclosed*: stopped naturally with / without closing the thought; *premature*: stopped with no boxed answer. Parentheses: paired difference; 95% CI in brackets; * significant; "–" not evaluated.

prompt format	model	cap	acc	solved-then-capped	loop	well-formed
EXACT	CE student	49.7	26.1	1.1	11.1	50.3
	A on-policy terminal	46.1 (−3.6)	24.7 (−1.4)	0.6 (−0.6)	11.9 (+0.8)	53.9 (+3.6)
	A, full-sequence loss	70.0 (+20.3*)	23.1 (−3.1)	2.5 (+1.4)	55.8 (+44.7*)	30.0 (−20.3*)
	B teacher-prefix terminal	50.6 (+0.8)	26.9 (+0.8)	1.7 (+0.6)	14.7 (+3.6)	49.4 (−0.8)
	B, full-sequence loss	47.2 (−2.5)	26.7 (+0.6)	1.1	10.0 (−1.1)	52.8 (+2.5)
	C replay only	49.7	26.7 (+0.6)	0.6 (−0.6)	13.1 (+1.9)	50.3
	D format-augmented replay	47.5 (−2.2)	26.4 (+0.3)	0.0 (−1.1*)	9.7 (−1.4)	52.5 (+2.2)
NO-SYSTEM	CE student	67.2	24.4	8.6	25.0	21.7
	A on-policy terminal	50.0 (−17.2*)	24.7 (+0.3)	1.4 (−7.2*)	12.8 (−12.2*)	48.3 (+26.7*)
	A, full-sequence loss	73.6 (+6.4)	23.9 (−0.6)	5.3 (−3.3)	60.3 (+35.3*)	25.8 (+4.2)
	B teacher-prefix terminal	50.3 (−16.9*)	27.2 (+2.8)	1.1 (−7.5*)	16.4 (−8.6*)	47.8 (+26.1*)
	B, full-sequence loss	51.7 (−15.6*)	26.1 (+1.7)	0.6 (−8.1*)	13.3 (−11.7*)	48.1 (+26.4*)
	C replay only	67.8 (+0.6)	25.8 (+1.4)	10.8 (+2.2)	24.2 (−0.8)	18.9 (−2.8)
	D format-augmented replay	50.8 (−16.4*)	28.1 (+3.6)	0.8 (−7.8*)	12.5 (−12.5*)	49.2 (+27.5*)

A17 Additional material

A17.1 Limitations in full

Three training domains (mathematics, science, code), about 30k examples each, with the code student from a different teacher and trace format; evaluation covers only reasoning benchmarks (MATH-500, AIME 2025, GPQA-Diamond, code problems), with no short-answer or non-reasoning benchmark such as GSM8K or MMLU-Pro; one trainer, one chat template family per base, one harness studied in depth. On mathematics the locally trained brittle students are all Qwen-derived and the robust ones Llama-3.1, Mistral and Qwen2.5-14B (on LiveCodeBench the Llama base student also meets the endpoint); the base-side displacement is a screen, not an explanation. The released students' prompt format histories are inferred from their recipes. The splice localizes cause (ii) at the two closing tokens on the 8B and 7B students only, and shows the closing conditional is sufficient to restore termination but not which circuit computes it; the optimizer result reuses checkpoints from separate runs of the same recipe and its mechanism is a hypothesis. The loop metric is a repetition statistic (agreement with independent detectors: Appendix A12.5) and "solved-then-capped" uses the last boxed answer. The dose sweep has one training run per fraction and seed, the splice one decoding seed, the on-policy arm one seed, the 32B replication arm D only, replication AIME cells n=120; stars are unadjusted for multiple comparisons, which matters for the small accuracy and premature-stop differences, not for the termination effects. The students' persistence under the training format after a wrong answer (cause (iii)) is decomposed in §5 and removed at decoding time, not in training; give-up sentences demonstrated from answer-less states collapse the student when shown alone and, anchored by replay, trade stop-after-wrong for accuracy on one training seed (Appendix A6); the arms meant to restore wrong endings restored few (Appendix A5). The format fix is shown to survive the ladder's added traces on one training seed; cause (i) was never crossed with the others, since the main recipe keeps the end-of-turn token (Appendix A10), and its prevalence is measured on two stacks' defaults and two corpora (Table 31), not on released training pipelines.

A17.2 Related work, in full

Prompt format, and what is read off it. That a model's output depends on surface choices in the prompt is well established: the order of in-context examples (Lu et al., 2022), separators and casing (Sclar et al., 2024), the template used for in-context learning (Voronov et al., 2024) and the paraphrase of an instruction (Mizrahi et al., 2024) all move accuracy, and perturbing the order of the answer options or the rule by which an answer is read off moves leaderboard positions by several places (Alzahrani et al., 2024); the recommended response has been to evaluate over many prompts rather than one and report the spread or its quantiles (Mizrahi et al., 2024; Maia Polo et al., 2024). A complementary line finds that the prompt's meaning does less work than its form:

models learn about as fast from misleading instructions as from meaningful ones (Webson & Pavlick, 2022), demonstrations contribute mainly the label space and the shape of the sequence rather than correct labels (Min et al., 2022), and putting a persona in the system turn does not make answers more accurate (Zheng et al., 2024). How much of the measured sensitivity belongs to the model is contested: Hua et al. (2025) report that replacing likelihood scoring and exact matching with a judge removes much of the variance across templates and stabilizes rankings. The same question is asked of evaluation protocols (Liang et al., 2023) and of harness implementations (Gao et al., 2023; Biderman et al., 2024). All of this varies a prompt that the model was not trained on in particular, and reads off whether a short answer is right. Our perturbation is defined by reference to the one prompt format the student's fine-tuning used, and our dependent variable is whether a ten-thousand-token generation ends at all—which accuracy cannot see, since it moves by tens of points here while accuracy moves by 0.6. An evaluation that averages accuracy over templates would therefore report this student as robust, and a headline termination number is interpretable only together with the format that produced it.

The template as an interface between training and serving. A mismatch between the template a model was tuned under and the one it is served with is a known hazard. Lyu et al. (2024a) show that fine-tuning on benign data can undo safety tuning, and that deliberately mismatching the two—tuning without a safety system prompt and adding it only at test time—retains much of the safety behaviour; Jiang et al. (2025) show that a user who violates the chat template at inference can bypass alignment across aligned models. Both establish that the template is load-bearing, and both ask what the model *says*. Robustness to it has been trained in by varying the prompt during fine-tuning, the move instruction tuning has made since Wei et al. (2022) and that prompt-agnostic fine-tuning makes explicitly (Wei et al., 2025). We add three things to this line: the outcome we measure is whether the model stops rather than what it says; the effect, for one family of students, is carried by the conditional distribution at a single token, which we show by intervening on that conditional; and the amount of format variety needed to remove it is small enough to state exactly—86 reformatted rows, 0.3% of the training set—and is tested on formats that no training arm saw.

Generation that does not end. That a language model can fail to terminate is an old result with a precise statement: Welleck et al. (2020a) define a decoding algorithm as inconsistent if it can return an infinite sequence to which the model assigns zero probability, prove that greedy search, beam search, top-k and nucleus sampling are all inconsistent, and repair it either by keeping the end-of-sequence probability reachable or by training a self-terminating model. That inconsistency is a property of the decoder, and every arm we compare shares one decoder, one sampler and one temperature, so it is common to all of them: it can raise the floor we measure against, but it cannot produce a difference between arms, which is the whole of what we report. In translation the same miscalibration is severe enough that the model's global optimum is often the empty string (Stahlberg & Byrne, 2019), which length corrections address directly (Murray & Chiang, 2018). Alongside this sits the degeneration literature: likelihood-trained models fall into repetition under likelihood-maximizing decoding (Holtzman et al., 2020), the loop is self-reinforcing (Xu et al., 2022) and admits a theoretical account (Fu et al., 2021), and it can be trained against (Welleck et al., 2020b). For reasoning models the failure is now measured: Ye et al. (2025a) report termination rate next to accuracy and find that reasoning models trail instruction-tuned ones on long procedural tasks mainly by not finishing within the budget, and Zhang et al. (2025b) find that a small model fine-tuned on long traces cannot locate the point at which to stop, and correct it with temperature scaling and length-regularized reinforcement learning. That work attributes non-termination to the decoding algorithm, the likelihood objective or the fine-tuned weights, all three of which our two conditions hold fixed. We add a fourth source: the same weights, decoded the same way, terminate or do not depending on eleven tokens of the prompt, and the behaviour can be moved either way by editing one coordinate of one conditional. Yang et al. (2026) independently report that the evaluation template shifts measured length and accuracy at fixed weights; we sharpen that observation to the serialization layer, to termination as the outcome, and to a localized cause.

Spending the token budget, and controlling it. Long chain-of-thought models spend more tokens than a problem needs (Chen et al., 2024; Sui et al., 2025), sometimes abandon good lines of attack (Wang et al., 2025), and trade length against reliability as reinforcement learning lengthens their traces (Yeo et al., 2025); distilled students in particular loop where their teachers do not (Pipis

et al., 2026). The remedies act at decoding time—forcing the budget (Muennighoff et al., 2025), exiting early (Yang et al., 2025b), stopping when the closing token becomes competitive (Song et al., 2025), or reading an answer-known signal off the hidden states (Zhang et al., 2025a)—or they train length behaviour in with reinforcement learning against a target supplied in the prompt (Aggarwal & Welleck, 2025); Snell et al. (2025) treats the budget itself as the quantity to allocate. All of these treat the failure to stop as a property of the model, to be corrected downstream of it. For students fine-tuned under a single prompt format it is substantially a property of the evaluation prompt, which sits upstream of every one of them: the same checkpoint, decoded identically, makes a well-formed stop in 84% of generations under one format and 54% under another, while accuracy moves 0.6 points.

Localizing an effect by intervening on it. Attributing a behaviour to a part of a network by changing that part and watching the behaviour follow is standard practice: causal mediation analysis (Vig et al., 2020), causal tracing, which corrupts a prompt's embeddings and restores individual hidden states until factual recall returns (Meng et al., 2022), and path patching (Wang et al., 2023; Goldowsky-Dill et al., 2023), whose metrics and pitfalls are by now documented (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Those interventions act on activations and localize in the network. Ours acts on the output distribution and localizes in behaviour: we replace the probability of one token at one class of decoding step with the value the same weights assign under the other prompt format, rescale the remaining mass so that no other preference changes, and let the model go on sampling—so what we read off is the whole generation rather than a change in one logit. Editing the output distribution is otherwise a decoding *method*, used to improve text rather than to test where an effect lives (Li et al., 2023; Chuang et al., 2024; Shi et al., 2024). We use it as a test, with sham splices on tokens that move more between the two formats, a reverse splice, and a gate restricting the intervention to the steps at which the training format would have closed. What it does not give is the circuit that computes the closing conditional, and on the 32B student the two formats leave that conditional identical, so the intervention has nothing to act on.

A17.3 Extended related work

Beyond the comparisons in the body: stopping is also controlled through trained positions and budgets (Han et al., 2025; Teknium et al., 2025), studied as a representation property of the end-of-sequence token (Newman et al., 2020), and distilled students are known to over-generate when trained on long traces (Li et al., 2025; Yu et al., 2025); none separates the causes by the branch of the identity they move.

What varies, and what is measured. Table 79 sets the closest lines of work against ours on six axes. The prompt-sensitivity literature varies a prompt the model was not specifically trained on and reads off a short answer: example order (Lu et al., 2022), separators and casing (Sclar et al., 2024), format under few-shot conditions (Voronov et al., 2024), instruction paraphrases (Mizrahi et al., 2024). Two results in that line are close to ours in spirit. Min et al. (2022) find that what demonstrations contribute is mostly the *format* of the sequence rather than the content of the labels, and Webson & Pavlick (2022) find that models learn about as fast from misleading instructions as from meaningful ones—in both cases the surface form is doing work that a reader would attribute to the semantics. Our result is the same shape one level down: the eleven tokens that the student conditions its stop decision on carry no task information at all, and removing them does not change what the student computes, only whether it ends.

Why a fine-tuned student and not a base model. Zhao et al. (2021) attribute few-shot instability to biases a *pretrained* model brings to the prompt. The brittleness here is absent from every base checkpoint we probed and appears after SFT (§4), so it is manufactured, not inherited; the closest general account is shortcut learning (Geirhos et al., 2020), with the particular shortcut being a once-per-example cue for a once-per-example decision.

Templates as an interface. Lyu et al. (2024a) show that a train/serve template mismatch changes post-SFT safety behaviour, and use it deliberately as a defence; Jiang et al. (2025) show that violating the chat template at inference defeats alignment across aligned models. Both establish that the template is a load-bearing interface. Neither measures termination, and neither localizes the effect

Table 79: Where this paper sits. Rows are the closest lines of prior work, grouped by theme; the last row is ours. *what is varied*: the quantity the study manipulates. *tuned on it?*: *yes* if the perturbation is defined by reference to a prompt format the model under test was itself fine-tuned or aligned under; *no* if it varies formats none of which the model was trained on in particular; *n/a* if no prompt format is varied. *what is measured*: the dependent variable. *stopping measured?*: whether a reported outcome is whether generation ends; *length only*, output length is reported but ending is not at issue; *loops only*, repetition is counted but not the failure to end. *causal intervention*: whether the mechanism is tested by manipulating the model or its decoding, and of what kind. *fix, and where tested*: the remedy and the conditions it is evaluated under. Qualitative positioning; we do not reproduce these papers' numbers. *RL*: reinforcement learning. *SFT*: supervised fine-tuning.

	what is varied	tuned on it?	what is measured	stopping measured?	causal intervention	fix, and where tested
Prompt format, and what is read off it						
Sclar et al. (2024)	separators, casing, spacing	no	spread of few-shot accuracy	no	no	report an interval
Mizrahi et al. (2024)	instruction paraphrases	no	accuracy and model rank	no	no	evaluate over many prompts
Alzahrani et al. (2024)	option order, answer scoring	no	leaderboard position	no	no	a sturdier scoring rule
Hua et al. (2025)	templates, and the scoring rule	no	variance of accuracy	no	no	score with a judge instead
Zheng et al. (2024)	persona in the system turn	no	factual accuracy	no	no	none; a negative result
The template as an interface between training and serving						
Lyu et al. (2024a)	tuning template vs. test template	yes	safety behaviour after tuning	no	no	mismatch the two on purpose
Jiang et al. (2025)	conformity to the chat template	yes	attack success rate	no	no	partial mitigations
Wei et al. (2025)	prompts seen while fine-tuning	yes	downstream accuracy	no	no	prompt-varied fine-tuning
Generation that does not end						
Welleck et al. (2020a)	the decoding algorithm	n/a	rate of non-termination	yes	no; a proof	consistent sampling; a self-terminating model
Stahlberg & Byrne (2019)	exact vs. beam search	n/a	the global optimum of the model	yes	exact search	a minimum-length constraint
Holtzman et al. (2020)	the decoding algorithm	n/a	repetition, distribution statistics	loops only	no	nucleus sampling
Zhang et al. (2025b)	the model, before vs. after SFT	no	length, repetition, accuracy	yes	no	temperature scaling; length-regularized RL
Pipis et al. (2026)	temperature, scale, distillation	n/a	rate of looping	loops only	on a synthetic task	temperature; insufficient
Ye et al. (2025a)	task, output budget, model	n/a	accuracy and termination rate	yes	no	none; a benchmark
Yang et al. (2026)	teacher vs. student EOS token; training and evaluation template	partly	response length, clipping, accuracy	length only	yes; in the objective	realign the stop action in the distillation objective; on-policy maths
Fu et al. (2026)	sampled-token supervision; special tokens	n/a	accuracy, degenerate output	loops only	yes; in the objective	mask special tokens from the signal; on-policy maths and agents
Spending the token budget, and controlling it						
Chen et al. (2024)	problem difficulty	n/a	tokens spent per correct answer	length only	no	train for efficiency
Muennighoff et al. (2025)	the thinking budget	n/a	length and accuracy	length only	closure forced at decoding	budget forcing
Aggarwal & Welleck (2025)	a length target in the prompt	n/a	length adherence, accuracy	length only	no	RL against the target
Song et al. (2025)	where the thought is cut	n/a	tokens and accuracy	yes	inject the closing token	stop on a log-probability margin
Localizing a behaviour by intervening on the model						
Meng et al. (2022)	hidden states, corrupted then restored	n/a	probability of the correct fact	no	activation patching	a weight edit
Wang et al. (2023)	activations along single edges	n/a	difference of two logits	no	path patching	n/a
this paper	the eleven-token system turn of the training format; six formats	yes	**whether generation ever ends**, over ∼10k tokens; accuracy unchanged	**yes; it is the outcome**	splice of one conditional, with sham, reverse and gated arms	86 reformatted rows, tested on formats held out of training; a fixed-point decode guard

to a conditional: our splice (§4) shows that for one family of students the whole effect can be carried by the probability of a single token.

Fixes. Prompt-varied fine-tuning (Wei et al., 2025) and the template diversity of instruction tuning (Wei et al., 2022) both add prompt variety during training for robustness of the answer. Our arm D is the same move applied to the prompt format rather than the instruction, and the finding is the dose: 86 rows (0.3% of the training set) recover termination under formats never trained on. Decode-side remedies—budget forcing (Muennighoff et al., 2025), early exit (Yang et al., 2025b)—and RL length control (Aggarwal & Welleck, 2025) act after the fact. Our decode guard is one such remedy, and budget forcing at the full budget does as well as it (§6): both supply the commitment the stuck student lacks, though the decode guard does not replace the data guard. Snell et al. (2025) makes the budget itself the object of optimization, which our result makes format-dependent.

Reasoning-trace corpora. The students here are trained the way open reasoning-trace distillations are (DeepSeek-AI, 2025; Du et al., 2025; Guha et al., 2026; Muennighoff et al., 2025; Ye et al., 2025b): one corpus, one prompt format, many epochs. That is exactly the condition under which we find the brittleness, which is why we think the phenomenon is common rather than particular to our runs—though, as §4 reports, the released students we tested do not all show it.

A17.4 EXAMPLES

A solved-then-capped generation of the CE student under the HISTORICAL format (AIME 2025 problem 1, reference answer 70, budget 30,000 tokens). The correct answer is written after 0.8% of the generation; after about 1,000 tokens of re-verification the model verbalizes its intention to stop for the remaining ∼28,700 tokens without ever emitting </think>:

```
... So both 21 and 49 are valid.

Thus sum = 21+49 = 70.

Answer: \boxed{70}.

We need to double-check no other b > 9 might satisfy the divisibility condition ...
[~1,000 tokens of re-verification]
... Now produce final answer.
Ok.
Now produce final answer.
Ok.
Now produce final answer.
Ok.
[repeated until the budget]
```

A second decoding seed of the same problem loops on a three-line period ("Thus final answer: \boxed{70}. / (Initialized.) / Ok."). Under the EXACT format the same checkpoint stops on this problem in every seed. The sequence shows that what fails is not the model's knowledge of the answer or of the fact that it should stop, but the emission of the tokens that end the thought and the turn.